



HAL
open science

How to deal with Discourse Markers : a prosodic, corpus-based, computational and experimental proposal

Saulo Mendes Santos

► **To cite this version:**

Saulo Mendes Santos. How to deal with Discourse Markers : a prosodic, corpus-based, computational and experimental proposal. Computation and Language [cs.CL]. Université Paris-Saclay; Universidade federal de Minas Gerais, 2024. English. NNT : 2024UPASG013 . tel-04594427

HAL Id: tel-04594427

<https://theses.hal.science/tel-04594427>

Submitted on 30 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

How to deal with Discourse Markers: a prosodic, corpus-based, computational and experimental proposal

*Comment aborder les Marqueurs Discursifs : une approche prosodique,
basée sur corpus, computationnelle et expérimentale*

Thèse de doctorat de l'université Paris-Saclay et de l'Universidade Federal de Minas Gerais

École doctorale n°580 Sciences et technologies de l'information et de la
communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique. Référent : Faculté des sciences d'Orsay

Thèse préparée en co-tutelle dans le **Laboratoire interdisciplinaire des sciences du
numérique (Université Paris-Saclay, CNRS)** et dans le **Laboratório de Estudos
Empíricos e Experimentais da Linguagem (Universidade Federal de Minas Gerais)**,
sous la direction d'**Albert RILLIARD**, Chargé de Recherche, et la co-direction de
Tommaso RASO, Professeur

Thèse soutenue à Paris-Saclay, le 26 février 2024, par

Saulo MENDES SANTOS

Composition du Jury

Membres du jury avec voix délibérative

Corinne FREDOUILLE Professeure, Université d'Avignon	Présidente
Plínio AMEIDA BARBOSA Professeur, Universidade Estadual de Campinas	Rapporteur & Examineur
Antonio ROMANO Professeur, Università degli Studi di Torino	Rapporteur & Examineur
Bruno ROCHA Professeur assistant, Univ. Fed. de Minas Gerais	Examineur
Alessandro PANUNZI Professeur agrégé, Università degli Studi di Firenze	Examineur

Titre : Comment aborder les Marqueurs Discursifs : une approche prosodique, basée sur corpus, computationnelle et expérimentale

Mots clés : Pragmatique ; Prosodie ; Structure Informationnelle ; Marqueurs Discursifs ; Traitement Automatique du Langage

Résumé : L'objectif principal de cette recherche est d'examiner les marqueurs discursifs (MD) dans les interactions vocales spontanées. Les MDs sont des lexèmes ou des expressions qui ont subi une pragmatization. Ils ne participent pas à la construction du sens, étant orientés vers la gestion des interactions parlées. Les fonctions proposées varient beaucoup selon les objectifs et la méthode d'analyse, et la plupart des études prennent le lexique comme point de départ. Cependant, une même fonction peut être réalisée à travers plusieurs lexèmes. Cette recherche adopte une approche différente. Le critère formel est la forme prosodique, considérée comme plus stable et révélatrice au regard des fonctionnalités des MDs. L'objectif est de comprendre les facteurs qui contribuent à la transmission des fonctions des MDs et comment la forme prosodique peut être utilisée pour prédire leurs fonctions respectives. Le cadre théorique utilisé est Language into Act Theory (L-Act – Cavalcante, 2020 ; Cresti, 2000 ; Moneglia & Raso, 2014). La L-Act considère les MDs comme un type particulier d'unité informationnelle (UI), correspondant à des unités prosodiques et soumise à des restrictions de distribution. Sur la base d'études antérieures (Cresti, 2000 ; Raso, 2014 ; Raso & Vieira, 2016 ; Raso & Ferrari, 2020), un cadre descriptif révisé des MDs est proposé avec les fonctions suivantes : Allocutive (cohésion sociale) ; Conative (indiquant une solution illocutoire) ; Expressive (surprise non illocutoire) ; Marquage (surligne); et Incipit (ouverture d'une TU ou d'un tour de parole). Des preuves statistiques et expérimentales soutenant la

proposition sont présentées. Un échantillon d'UIs contenant des candidats MDs a été extrait du corpus C-ORAL-BRASIL I (Raso & Mello, 2012). Les données ont été classées selon les cinq classes fonctionnelles proposées. 30 descripteurs acoustico-prosodiques ont été estimés, incluant la forme intonative, l'allongement syllabique et l'intensité relative. Différents modèles de classification ont été entraînés et évalués en validation croisée. Le meilleur modèle de classification atteint une précision de 78 % pour cinq classes. Ces résultats montrent que, malgré la grande variabilité des contextes, des locuteurs et des styles d'énoncé, il est possible de parvenir à une classification fonctionnelle raisonnable grâce à la forme prosodique des MDs. Enfin, une évaluation perceptive étudie la capacité des participants à identifier les classes de MDs qui se produisent en position initiale, uniquement sur la base d'indices prosodiques. 25 énoncés contenant trois fonctions différentes (ALL, CNT et INP) et sept lexèmes différents ont été sélectionnés. Ces exemples ont été reproduits en chambre sourde par un locuteur expérimenté et ces originaux manipulés pour porter les formes prototypiques des fonctions cibles. Ces stimuli ont été présentés à 120 participants. Les résultats montrent que tous les facteurs contrôlés jouent un rôle significatif : le contexte original, la forme prosodique et le lexique. Cependant, seule la forme prosodique indique, de manière moins variable et moins soumise à l'interprétation des phrases, la fonction visée.

Title : How to deal with Discourse Markers: a prosodic, corpus-based, computational and experimental proposal

Keywords : Pragmatics; Prosody; Information Structure; Discourse Markers; Natural Language Processing

Abstract : The primary focus of this research is to examine Discourse Markers (DMs) in spontaneous spoken interactions. DMs are lexemes or small expressions that underwent pragmaticalization. They do not participate in constructing the meaning, being directed at managing spoken interactions. Proposed DM functions vary greatly depending on goals and analytical methods, with most studies taking the lexicon as a departing point. However, the same DM function can be accomplished through various lexemes. This research takes a different approach. The formal criterion is the prosodic form, which is deemed more stable and revealing regarding the functionalities of DMs. Thus, the objective is to comprehend the factors contributing to expressing DMs' functions and understand how their prosodic form can predict their respective functions. The theoretical framework utilized for this investigation is the Language into Act Theory (L-Act – Cavalcante, 2020; Cresti, 2000; Moneglia & Raso, 2014). The L-Act views DMs as a special type of Information Unit (IU), which are conveyed by prosodic units and have distributional constraints. Based on previous studies (Cresti, 2000; Raso, 2014; Raso & Vieira, 2016; Raso & Ferrari, 2020), a reviewed DM framework is proposed: the Allocutive (social cohesion); the Conative (pointing to an illocutionary solution); the Expressive (enacting non-illocutionary surprise); the Highlighter (highlighting); and the Incipit (opening a TU or a turn). Statistical and experimental evidence supporting the proposal is presented. An IU sample containing DM candidates was extracted from the C-ORAL-BRASIL I corpus (Raso & Mello, 2012).

The data were categorized into the five proposed classes. 30 prosodic-acoustic descriptors were estimated, including intonational shape, syllable lengthening, and relative intensity. Different classification models were trained and evaluated on a cross-validation set. The best classification model achieved an accuracy score of 78% for these five categories. The more relevant features for distinguishing each class from others are described. The results show that despite the large variability of contexts, speakers, and speaking styles, achieving a reasonable classification through the prosodic form is possible. Additionally, an experiment evaluated participants' ability to recognize DM classes occurring in the initial position based on prosodic cues. 25 utterances containing three different DM functions (ALL, CNT, and INP) and seven different lexemes were selected. A trained speaker reproduced instances in a sound-processed room, and the original DMs were manipulated to match the prototypical forms of all functional targets. The stimuli were presented to 120 participants, whose task was to identify the function. Results show that all controlled factors were relevant: the original context, the prosodic form, and the lexicon. However, the prosodic form is less variable and less submitted to the sentence's interpretation to cue functionality.

Título : Como lidar com os Marcadores Discursivos : uma abordagem prosódica, baseada em corpus, computacional e experimental

Palavras-chave : Pragmática; Prosódia; Estrutura Informacional; Marcadores Discursivos; Processamento da Linguagem Natural

Resumo : O foco principal desta pesquisa é examinar Marcadores de Discurso (MDs) em interações de fala espontânea. DMs são lexemas ou expressões que passaram por pragmaticalização. Eles não participam da construção do significado, sendo direcionados ao gerenciamento das interações faladas. As funções propostas variam muito dependendo dos objetivos e do método analítico, e a maioria dos estudos toma o léxico como ponto de partida. Contudo, uma mesma função pode ser realizada através de vários lexemas. Esta pesquisa adota uma abordagem diferente. O critério formal é a forma prosódica, considerada mais estável e reveladora no que diz respeito às funcionalidades dos DMs. O objetivo é compreender os fatores que contribuem para a veiculação dos DMs e como a forma prosódica pode ser utilizada para prever as respectivas funções. O referencial teórico utilizado é a Language into Act (L-Act – Cavalcante, 2020; Cresti, 2000; Moneglia & Raso, 2014). A L-Act vê os DMs como um tipo especial de Unidade Informacional (UI), que é veiculada por unidades prosódicas e possui restrições distribucionais. Com base em estudos anteriores (Cresti, 2000; Raso, 2014; Raso & Vieira, 2016; Raso & Ferrari, 2020), propõe-se um quadro de MDs revisto: o Alocutivo (coesão social); o Conativo (apontando para uma solução ilocucionária); o Expressivo (representando surpresa não ilocucionária); o Marcador (destaque); e o Incipitário (abrir uma TU ou um turno). São apresentadas evidências estatísticas e experimentais que apoiam a proposta.

Uma amostra de UIs contendo candidatos a DM foi extraída do C-ORAL-BRASIL I (Raso & Mello, 2012). Os dados foram categorizados nas cinco classes propostas. Foram estimados 30 descritores prosódico-acústicos, incluindo forma entoacional, alongamento silábico e intensidade relativa. Diferentes modelos de classificação foram treinados e avaliados em validação cruzada. O melhor modelo de classificação obteve uma acurácia de 78%. Os resultados mostram que, apesar da grande variabilidade de contextos, falantes e estilos de elocução, é possível conseguir uma classificação razoável por meio da forma prosódica. Além disso, foi realizado um experimento para avaliar a capacidade dos participantes de reconhecer classes de DM que ocorrem na posição inicial com base apenas em pistas prosódicas. Foram selecionados 25 enunciados contendo três funções diferentes (ALL, CNT e INP) e sete lexemas diferentes. As instâncias foram registradas em um ambiente controlado e os originais foram manipulados para corresponder às formas prototípicas de outros DMs alvo. Os estímulos foram apresentados a 120 participantes. Os resultados mostram que todos os fatores controlados desempenham um papel relevante: o contexto original, a forma prosódica e o léxico. No entanto, apenas a forma prosódica é menos variável.

RESUMÉ SUBSTANTIEL EN FRANÇAIS

L'objectif principal de cette recherche était d'étudier les Marqueurs discursifs (MD) dans les corpus de parole spontanée. Dans la littérature, le point de départ de l'étude des MD est généralement le lexique ; on peut choisir, par exemple, le lexème GENRE et, sur la base de critères syntaxiques, distributionnels ou contextuels, étudier les fonctions que cet élément peut assumer. Cette procédure entraîne des problèmes cruciaux. Les modèles basés sur le lexique n'ont pas permis de définir, d'expliquer et de prédire clairement les fonctions des MD. La réponse à ce problème a été ancrée dans la prosodie des Marqueurs discursifs. À cette fin, j'ai expliqué l'importance de l'analyse prosodique de la parole et comment le discours est organisé par la théorie sur laquelle se base cette recherche, la Language into Act Theory (L-Act – Cavalcante, 2020 ; Cresti, 2000 ; Moneglia & Raso, 2014). J'ai discuté de la façon dont les MD peuvent être définis de manière à permettre leur prédiction et comment on peut discriminer leurs fonctions sur la base de leurs formes prosodiques (Cresti, 2000 ; Frosali, 2008 ; Raso et al., 2022 ; Raso & Vieira, 2016). J'ai passé en revue les travaux antérieurs sur les fonctions et les formes théoriques des MD et présenté une proposition contenant cinq macro-fonctions des MD. Les fondements d'un modèle visant à expliquer la projection des formes prosodiques sur les fonctions théoriques proposées des MD ont également été esquissés, ainsi qu'une exploration de la mesure dans laquelle cette projection est possible. De plus, la validation de la proposition d'un point de vue statistique et expérimental a été abordée.

Une gamme variée de corrélats acoustiques de hauteur de voix, d'intensité et de rythme de la parole a été estimée pour modéliser les MD. Un problème fréquent lors de l'analyse des données concerne l'estimation et le traçage de la courbe de la fréquence fondamentale (f_0), le corrélat de la hauteur de voix. Cela est particulièrement vrai lorsqu'il s'agit de données de parole spontanée. Enregistrées en dehors d'environnements acoustiquement isolés et contrôlés, le signal de parole peut être rapidement dégradé par la présence de réverbération, de climatiseurs, de ventilateurs, de bruits de moteurs de voiture ou d'autres phénomènes similaires. Cela peut conduire à ce

que les Algorithmes de Détection de la Hauteur de Voix (en anglais, Pitch Detection Algorithm – PDA) estiment de manière inexacte les valeurs de f_0 . Au cours des 30 dernières années, plusieurs méthodes et algorithmes ont été proposés pour faire face à de tels problèmes. Chaque méthode présente différentes sensibilités aux diverses conditions dégradant le signal de parole. Un autre problème est que les données de parole enregistrées dans des paramètres non contrôlés peuvent nécessiter que les chercheurs révisent manuellement les données de f_0 et ajustent divers paramètres fournis par les algorithmes PDA. Étant donné que chaque environnement d'enregistrement présente des conditions acoustiques différentes, on peut supposer que chaque fichier audio nécessitera une paramétrisation différente pour obtenir l'estimation et le traçage de f_0 optimaux. Assurer la reproductibilité du travail nécessiterait de garder une trace de toutes les modifications apportées aux données de f_0 et de toutes les paramétrisations appliquées à chaque fichier audio analysé. Pour rationaliser cette tâche potentiellement lourde, chronophage et sujette aux erreurs, la solution proposée améliore le traçage de f_0 en comparant la sortie de plusieurs algorithmes PDA avec un modèle de décision voisé/non-voisé.

Un autre aspect important de ce travail a été les efforts déployés pour une validation des fonctions proposées des MD d'un point de vue perceptuel. Par exemple, l'approche basée sur le corpus présentée par Lee et al. (2020), qui ont extrait un large ensemble de MD de différents corpus et les ont classés selon leurs fonctions, a permis aux auteurs d'observer des corrélations entre les fonctions des MD et leurs caractéristiques prosodiques, dans deux langues (anglais et français) et plusieurs styles de parole. Pendant ce temps, la validation perceptuelle de l'importance des indices prosodiques pour les diverses fonctions des MD est encore rare dans la littérature : le travail proposé dans Didirková et al. (2019) constituant une étape importante pour la validation de telles relations en français. Dans cette recherche, une méthodologie pour la validation de la relation entre les fonctions des MD et leurs formes prosodiques a été proposée.

Les Marqueurs de Discours ont été définis dans ce travail comme des unités informatives ayant des fonctions interactionnelles.

En tant que tels, les MD sont véhiculés à travers des unités prosodiques, ils ont des macro-fonctions spécifiques qui sont transmises par (ou associées à) une forme prosodique, et ils ont des préférences distributionnelles. Les MD ne sont pas compositionnels par rapport aux structures qu'ils parasitent. Au lieu de cela, ils visent à réguler les aspects interactionnels du discours. Ils peuvent promouvoir la cohésion sociale (ALL), attirer l'attention du destinataire sur une solution illocutoire (CNT), exprimer la surprise sans force illocutoire (EXP), mettre en évidence un contenu précédent (EVD/HGL), ou simplement commencer l'énoncé (INP). Chaque fonction proposée peut être remplie avec une gamme variée de lexèmes ou d'expressions courtes. Il a été démontré que le lexique est variable tandis que la forme prosodique peut rendre compte de la reconnaissance des fonctions proposées avec de bonnes performances dans une tâche de classification. Il a également été démontré que, globalement, la prosodie a des effets positifs sur la reconnaissance des fonctions des MD. Néanmoins, il a également été démontré que le lexique joue également un rôle important dans l'interprétation des fonctions des MDs.

Un modèle de classification a été présenté avec les caractéristiques les plus pertinentes pour la distinction de chaque classe de MD contre les autres. Il est possible de dire que le modèle de classification présente de bonnes performances (scores de précision variant entre 68% et 78% pour cinq classes). Ce modèle ne présente pas le même niveau de précision que ceux présentés dans Gobbo (2019) – environ 80%. Cependant, la proposition actuelle (et le modèle respectif) est plus complexe : la tâche a été réalisée non pas avec trois mais cinq classes de MD. De plus, le modèle actuel tient compte des observations précédemment non classées qui ont été laissées de côté soit parce qu'elles étaient ambiguës, soit parce qu'elles ne correspondaient à aucune classe existante. Enfin, le modèle a été évalué avec des techniques plus robustes, et ses performances posent moins de questions écologiques. Une autre observation intéressante concerne les caractéristiques les plus fréquemment choisies par les modèles « un contre le reste ». Dans la plupart des cas, les caractéristiques impliquant la fréquence fondamentale étaient importantes. Les exceptions sont la classe ALL, qui sélectionne la durée

(plus longue), et INP, qui sélectionne l'intensité (plus élevée) et la durée (plus courte). Les caractéristiques d'alignement se sont également avérées pertinentes pour les distinctions.

Une approche intéressante pour un modèle plus compréhensible pourrait être d'avoir des modèles dédiés à des caractéristiques spécifiques. Un ensemble pourrait être construit qui regroupe des modèles, chacun dédié à une facette des MDs. Par exemple, un modèle pourrait voter exclusivement sur la base des caractéristiques prosodiques. Un autre pourrait être en charge des aspects liés à la distribution du MD dans le schéma. Non seulement une caractéristique catégorielle indiquant la position (initiale/médiane/finale) pourrait s'avérer utile, mais également des caractéristiques plus fines pourraient être testées qui reflètent la distance relative du MD par rapport à l'unité illocutoire (à la fois en unité de temps relative et en nombre d'unités informatives), ainsi que les unités informatives voisines. Un autre modèle pourrait être responsable de juger la classe basée sur le remplissage lexical du MD. Cela pourrait être réalisé en utilisant des *sentences embeddings* (*sentence transformers* – Reimers & Gurevych, 2020) comme entrée. Cela empêcherait, par exemple, que les prénoms (fréquemment utilisés dans CNT et INP) soient traités comme des catégories très divergentes, comme un simple encodage catégoriel du texte du MD.

De plus, certaines conclusions importantes ont été tirées des tests perceptuels. Premièrement, certains lexèmes ont reçu de fortes associations avec certaines fonctions. Ces associations semblent être favorisées ou défavorisées par la prosodie (et par l'illocution). Par exemple, 'não' (non) a une fonction conclusive, mais la perception de ce lexème comme conclusif devient plus saillante lorsque la forme prosodique est descendante et moins saillante lorsqu'elle est ascendante. Mais l'attribution catégorielle quasi systématique de fonctionnalité par les participants du test à certains lexèmes (par exemple, 'uai' ou 'eh') ne doit pas être exagérée : d'abord, elle peut varier beaucoup selon l'interprétation contextuelle du lexème (aucun lexème n'a une interprétation fixe à partir de la forme écrite, dans toutes les énonciations), et deuxièmement – si les participants se fient à la signification de base des lexèmes, les observations du corpus

montrent que ces lexèmes sont utilisés dans une variété de contextes. Il se peut que les participants aient du mal avec la nature désémantisée des MD dans ce cas. D'autres travaux seront nécessaires pour offrir des protocoles expérimentaux capables de faire face à cette limitation (un protocole d'association, comme dans Shochi et al., 2020, peut s'avérer intéressant). Cependant, le lexique offre un large éventail de possibilités, surtout depuis que les noms propres peuvent être utilisés comme MD dans CNT et ALL. De plus, le lexique est pluri-fonctionnel ; la signification des lexèmes dépend fortement du contexte. Cela rend une classification fonctionnelle à partir du lexique compliquée. En outre, le lexique est très variable diatopiquement, diaphasiquement et diachroniquement, rendant une classification encore plus compliquée. Deuxièmement, le type d'illocution et son contenu sémantique se sont avérés avoir un effet sur l'interprétation globale de l'énoncé. Bien sûr, ce qui était demandé était l'interprétation du MD, qui est intégré dans l'énoncé, mais toute la structure produit une signification globale. Il est raisonnable de penser que le rôle fonctionnel du MD est interprété de manière holistique au sein de l'énoncé. Par conséquent, si l'illocution porte (ou est interprétée comme) une surprise illocutoire, une conclusion, ou autre chose, cela affecte l'interprétation du MD par le participant. Cependant, la fonction du MD est, dans une large mesure, indépendante de l'illocution. Il semble sensé de penser que l'illocution impose certaines contraintes combinatoires sur le MD, mais il doit y avoir un certain degré de liberté. De plus, il existe également de nombreuses catégories d'illocution (dont la plupart ont encore besoin de descriptions plus approfondies), et ce facteur n'a pas pu être contrôlé. Ce qui a été pris en compte était un facteur de classe DM original simplifié (qui devrait résumer toutes les caractéristiques naturelles de l'énoncé original).

Cette recherche a identifié cinq formes prosodiques qui semblent fonctionnellement cohérentes et suffisantes pour couvrir toutes les fonctions des MD. La prosodie peut varier diatopiquement et diaphasiquement (peut-être aussi diachroniquement) en fonction des paramètres attitudeux : une intensité plus élevée ou plus basse, une plage de f_0 , un taux d'articulation peuvent indubitablement dépendre des caractéristiques démographiques des participants (genre, âge, niveau socio-culturel, et autres) et de la situation (les gens

ajustent leur attitude en fonction de la situation de communication). Mais la forme prosodique (mouvement et alignement, avant tout) reste, dans une large mesure, constante. Cette constance est exactement ce qui a permis d'obtenir de bons scores sur les tâches de classification.

Sur la base de ces considérations, la conclusion provisoire tirée des expériences est que pour catégoriser les MD, étant donné que de nombreux facteurs influencent leur interprétation, il faut commencer par le facteur le moins variable – la forme prosodique, sachant que d'autres facteurs peuvent modifier l'interprétation de base de la forme et peuvent même la modifier beaucoup. On ne devrait pas, de manière avisée, commencer par les facteurs qui varient le plus, tels que le lexique, les attitudes, ou les types d'illocutions. Ce sont tous des facteurs avec des degrés importants de variation qui ne permettent pas une organisation initiale. Si nous, à titre d'exemple, commençons par le lexique, nous arriverons à la conclusion que le même lexème peut accomplir des fonctions essentiellement différentes et que la même fonction peut être accomplie par des lexèmes entièrement différents, disons, un nom propre et un verbe.

Comme il a été argumenté, les MDs proposés basés sur la prosodie sont des macro-fonctions qui peuvent prendre des sous-fonctions plus spécifiques en fonction du contexte. Cependant, les sous-fonctions sont cohérentes avec les macro-fonctions. Par exemple, CNT est censé pointer vers la solution illocutoire, c'est-à-dire vers l'intention du locuteur. Si le locuteur dit quelque chose, puis s'interrompt pour introduire une nouvelle planification, cette réparation peut être introduite par un CNT (si elle pointe vers une conclusion) ou par un INP, si le locuteur veut, par exemple, marquer un fort contraste avec l'idéation interrompue.

Enfin, la conception des expériences a montré certaines limitations : la formulation des questions, la nature des données, la définition des catégories possibles utilisées pour répondre sont notablement complexes et peuvent ne pas permettre une compréhension aisée par certains participants. Les locuteurs naïfs ne sont pas enseignés pendant la scolarité à identifier la prosodie ou les

MD de la même manière qu'ils sont enseignés à interpréter un lexème comme 'uai' en tant qu'interjection pouvant exprimer la surprise. Le biais écrit, hérité du système éducatif, a peut-être joué un rôle significatif dans nos résultats. Une réflexion est donc nécessaire pour aider à concevoir d'autres expériences qui prennent en compte les problèmes observés pour les tâches métalinguistiques présentées aux participants naïfs. Une idée possible est de présenter uniquement des exemples naturels, sans manipulations et décontextualisations, et de demander aux participants d'identifier la fonction.

To my dear mother, who gave a world to/for me.

ACKNOWLEDGMENTS

This is actually the last bit written about this thesis. And possibly the most important one. Because without you all, none of this would have been possible.

So, I'd like to begin with expressing infinite gratitude to my supervisors. Tommaso and Albert, you gave your ideas, your time (a lot of your time), unflagging support, guidance, and exemplar mentorship throughout this journey. I am eternally grateful for your insights and constructive feedback, which have been crucial in the shaping of this thesis. There is, of course, that Neapolitan and French cultural input impossible not to mention! Working under your supervision has been an incredible privilege, and I want to take a moment to acknowledge the impact it has had on my academic and personal growth. Your expertise in Pragmatics, Prosody, Psycholinguistics, Statistics has been a guiding light, helping me navigate the many complexities of research and academia. Your dedication to excellence and your commitment to fostering a collaborative learning environment have made a significant difference in my experience as a doctoral candidate. Beyond the technical aspects of my research, I appreciate the personal support and encouragement you've offered during this process. I am grateful for the opportunities both of you have provided, from networking within the academic community to the chance to collaborate on meaningful projects (the LBASS courses, the study groups on programming for linguist, and other projects within the LISN and the LEEL). I hope one day I'll be able to give your work back to society, with the same quality and dedication that you put into all this work. Grazie mille, merci infiniment !

I extend my heartfelt thanks to the rapporteurs of this thesis, Plínio and Antonio, and to the other members of my thesis committee, Alessandro, Bruno, Corinne, Heliana, and Martine, for their thoughtful input, constructive criticism, and valuable suggestions that greatly enhanced the quality of this work.

I extend a special note of gratitude to Professors Heloísa Penna, Cécile Balkanski, Hélène Bonneau, and Anne Lacheret, who generously

provided me with the invaluable opportunity to gain teaching experience during my doctoral period. Their mentorship and guidance not only enhanced my pedagogical skills but also enriched my overall academic experience.

Special thanks are also due to the Initiative d'Excellence (Action Doctorale Internationale / 2020) and to the PROEX/CAPES program for their financial support, which enabled me to carry out tasks, experiments, attend conferences, and access crucial resources vital for the completion of this research. Here, I extend my appreciation to the Federal University of Minas Gerais and to the Paris-Saclay University, two publicly funded excellent institutions, for their steadfast commitment to academic excellence and the pursuit of knowledge, even in the most difficult and darkest times of recent political events. The resources, infrastructure, and opportunities provided by these institutions have played a pivotal role in shaping my doctoral experience. The accessibility to facilities, libraries, and computation servers were instrumental in conducting this research. I am very grateful for the dedication of the faculty staff and administrators who work tirelessly to create an environment that promotes learning and research. This acknowledgment is a tribute to the enduring impact of publicly funded universities in fostering education, research, and the development of future students and scholars walking the same path I walked. The support provided by the Federal University of Minas Gerais and the Paris-Saclay University has been invaluable. I am very proud of having been a part of these institutions.

People I encountered during this journey encouraging my work: Professors Lucia, Giulia, Heliana, Bruno, João, Marc. To Oliver, my predecessor in studying DMs, to whom I owe so much. Camila, Marianna, Gabriela, João Victor, Rémi, Simon. I would like to acknowledge the assistance and support of my colleagues and fellow researchers at UFMG and UPSaclay. Marc once more, thank you a lot for the ideas and collaboration in the ML models! João, thank you for your input and for being the voice of the experiments run for this thesis! The shared insights, and collaborative spirit of ya'll have been sheer inspiration!

A big thank you to the French-Brazilian infantry: Ernandes, João,

Vanessa; everyone from the *transtornos in Paris* friend group: Isabela, Luiza, José, Anaïs, Maria, Kevin, Valentin.

To my mum, Lourdinha, dad, Gelito, and sis', Paula! Words, especially written in English, cannot express my gratitude! Obrigado por todo o suporte emocional, financeiro, psicológico! Sem vocês, nada disso faria nenhum sentido. Aos primos queridos e àqueles mais próximos, Aline, Marina, Bidu! S2!

Finally, I want to express my deepest gratitude to my partner, Mika, for his love, patience, encouragement, and understanding throughout this voyage's waving ups and downs. Your support has been the guiding stars in the sky, the astrolabe in the ship, the anchor that kept me grounded in the ports along the way, the lighthouse in the dark seaside off the shore, the safe harbor to come to. Thank you!

This thesis is the result of the collective efforts, support, and inspiration from all those mentioned above. I am truly grateful for the privilege of undertaking this research and the opportunity to contribute, to the best of my ability, to the linguistic knowledge. Everything right is to be partaken with my mentors; errors are from my stubborn self.

*The entire universe is
perfused with signs, if it is
not composed exclusively of
signs.*

(Charles Sanders Peirce)

*I have never doubted the
truth of signs, Adso; they are
the only things man has
with which to orient himself
in the world.*

(The Name of the Rose,
Umberto Eco)

Science is the belief in the ignorance of experts.

(Richard Feynman)

This research was conducted with funding from IDEX/Paris-Saclay (ADI/2020 – November 2020 to October 2022) and CAPES (March 2023 to February 2024).

LIST OF FIGURES

Figure 1 - Prosodic form of illocutions in Audio 4.....	50
Figure 2 - Type 1 TOP	55
Figure 3 - Type 2 TOP	56
Figure 4 - Type 3 TOP	57
Figure 5 - APC	59
Figure 6 - APT	60
Figure 7 - PAR	62
Figure 8 - INT	63
Figure 9 - Form of high INP	83
Figure 10 - Form of flat INP	84
Figure 11 - Form of CNT	85
Figure 12 - Form of ALL	87
Figure 13 - Form of EXP	88
Figure 14 - Form of EVD	89
Figure 15 - Distribution of lexemes/small expressions.....	107
Figure 16 - Correspondence between IPA and ASCII characters.....	110
Figure 17 - Illustration of an annotated file	111
Figure 18 - Proportional difference.....	113
Figure 19 - Flowchart of the VD CNN model architecture	134

Figure 20 - Effect of SNR (all noises mixed) on the accuracy of the VD by the 11 PDAs and the three models.	137
Figure 21 - Accuracy per SNR level and noise type (plots), for our three models (all, f0, MFCC) and the tested PDAs	138
Figure 22 - F0 trellis.....	144
Figure 23 - Raw F0 estimations of six different PDAs for audio file bfamcv03_202.....	146
Figure 24 - Viterbi Path for audio file bfamcv03_202.....	148
Figure 25 - Predictions of the best voicing decision model for audio file bfamcv03_202.....	148
Figure 26 - Distribution of features of intensity by class of DM	153
Figure 27 - Correlation between features of intensity	157
Figure 28 - Distribution of the features of duration	158
Figure 29 - Correlation between features of duration	160
Figure 30 - Distribution of the features of f0.....	162
Figure 31 - Significant differences between pairs of DMs by features of f0	162
Figure 32 - Correlation between features of f0	165
Figure 33 - Distribution of features of f0 variation	167
Figure 34 - Correlation between features of f0 variation.....	170
Figure 35 - Distribution of the features of alignment.....	172
Figure 36 - Correlation between features of alignment	176
Figure 37 - Distribution of features of f0 curve	178

Figure 38 - Fitted curves by DM function using a cubic function	180
Figure 39 - Correlation between coefficients of the fitted curves.....	183
Figure 40 - SofImpute (left) and Iterative Imputer (right) results for file bfamcv07_114.....	188
Figure 41 - SofImpute (left) and Iterative Imputer (right) results for file bfamcv22_127.....	189
Figure 42 - F0 curve of six selected PDAs smoothed by Viterbi Algorithm	191
Figure 43 - Fitted curve vs original data of the quadratic function (Audio file bfamcv11_2)	193
Figure 44 - Fitted curve vs original data of the cubic function (Audio file bfamcv11_2)	194
Figure 45 - Fitted curve vs original data of the quartic function (Audio file bfamcv11_2)	194
Figure 46 - Prototypical f0 curves of each DM class using the cubic function	195
Figure 47 - Confusion matrix for an LDA model using coefficients of the cubic function	196
Figure 48 - Estimation of MSE (C_p) for the 60 best combinations of features	210
Figure 49 - Number of times each feature was selected among best combinations (%)	211
Figure 50 - Overall accuracy and max accuracy score as a function of number of features.....	212
Figure 51 - LDA plot for 15 features.....	214

Figure 52 - Cp statistic (ALL vs OTHERS).....	218
Figure 53 - Most selected features (ALL vs OTHERS).....	219
Figure 54 - Decision Tree plot (ALL vs OTHERS).....	221
Figure 55 - Cp statistic (CNT vs OTHERS).....	222
Figure 56 - Most selected features (CNT vs OTHERS).....	223
Figure 57 - Decision Tree plot (CNT vs OTHERS).....	225
Figure 58 - Cp statistic (EVD vs OTHERS).....	226
Figure 59 - Most selected features (EVD vs OTHERS).....	227
Figure 60 - Decision Tree plot (EVD vs OTHERS).....	229
Figure 61 - Cp statistic (EXP vs OTHERS).....	230
Figure 62 - Most selected features (EXP vs OTHERS).....	231
Figure 63 - Decision Tree plot (EXP vs OTHERS).....	233
Figure 64 - Cp statistic (INP vs OTHERS).....	234
Figure 65 - Most selected features (INP vs OTHERS).....	235
Figure 66 - Decision Tree (INP vs OTHERS).....	236
Figure 67 - F1-score and accuracy score as a function of number of features resulting from the SelectKBest algorithm using a stratified 10-fold cross- validation set.....	237
Figure 68 - Visual representation of Stratified k-fold Cross-validation	240
Figure 69 - Confusion Matrix - LOOCV - Undersampling	241
Figure 70 - Confusion Matrix - LOOCV - Oversampling.....	242

Figure 71 - Example of manipulations of the three utterances used in the discrimination task	249
Figure 72 - Manipulated DMs of the discrimination task	251
Figure 73 - Proportion of Match answers fitted by the binomial regression as a function of the type of Pair	260
Figure 74 - Proportion of the (CON, SUR, STA) answers for each level of the stimuli's presentation Modality	268
Figure 75 - Proportion of the (CON, SUR, STA) answers for each level of Lexeme used for the stimuli.....	270
Figure 76 - Proportion of the (CON, SUR, STA) answer for each level of the functional Class of the stimuli.....	272
Figure 77 - Interaction between modality and lexeme	274
Figure 78 - Proportion of the (CON, SUR, STA) answers for each level of the presentation Modality of the stimuli for each functional Class	276
Figure 79 - Proportion of the (CON, SUR, STA) answers for each level Lexeme and functional Class	277
Figure 80 - Proportion of the (CON, SUR, STA) answers for each level of Lexeme, functional Class, and Presentation Modality.....	279

LIST OF TABLES

Table 1 - Pragmatic and cognitive parameters	52
Table 2 - Synthetic table of the textual IUs assumed by the L-Act, their functions and main references	66
Table 3 - Synthetic table of tags given to other prosodic units	68
Table 4 - DM summary table.....	90
Table 5 - Informal subcorpus (C-ORAL-BRASIL I).....	94
Table 6 - Subcorpora and domains of use of the C-ORAL-BRASIL II	94
Table 7 - Kappa values for the realistic agreement rate before segmentation validation.....	99
Table 8 - DM tokens in the BP minicorpus (Gobbo, 2019)	102
Table 9 - Discarded tokens by criterion	103
Table 10 - DM distribution in the revised sample.....	104
Table 11 - Final sample used for the classification task.....	105
Table 12 - Lexical frequency by DM class.....	105
Table 13 - Speech corpora: name, language (Lang: English or Brazilian Portuguese, BP), number of speakers (Spk: Fe- male/Male), total duration (Dur, in minutes), proportion used for training and testing (Tr/Te, if applicable), and reference (Ref).....	128
Table 14 - Types of noise and source.....	130
Table 15 - List of the PDA tested in this study, with general characteristics	131

Table 16 - Performance of tested models.....	133
Table 18 - Hyperparameters, search spaces and selected parameter for the VD decision CNN model	135
Table 19 - Global Accuracy (Glob. acc.) observed on the test set for each PDA and Model: mean (standard deviation), all SNR and noise type mixed; Accuracy of these systems on the Clean part of the test set only (Clean Ac.); Accuracy estimated on the Unseen data	136
Table 20 - Significant differences between pairs of DMs by feature of intensity.....	154
Table 21 - Statistical summary of the features of intensity.....	155
Table 22 - Significant differences between pairs of DMs by features of duration.....	159
Table 23 - Statistical summary of the features of duration.....	159
Table 24 - Statistical summary of the features of f0.....	163
Table 25 - Significant differences between pairs of DMs by features of f0 variation	168
Table 26 - Summary statistics of the features of f0 variation.....	168
Table 27 - Significant differences between pairs of DMs by features of alignment	173
Table 28 - Summary statistics of the features of alignment.....	174
Table 29 - Significant differences between pairs of DMs by f0 curve coefficients.....	179
Table 30 - Statistics summary of the features of f0 curve	181
Table 31 - Classification performance based on f0 curve coefficients	193

Table 32 - Number of observations and proportions per DM class.....	198
Table 33 - Classification report of different classifiers	205
Table 34 - Coefficients of discriminative functions.....	214
Table 35 - Confusion matrix - Overall model	216
Table 36 - Performance metrics by class for the overall model	216
Table 37 - Model fit (ALL vs OTHERS).....	222
Table 38 - Model fit (CNT vs OTHERS).....	225
Table 39 - Model fit (EVD vs OTHERS).....	229
Table 40 - Model fit (EXP vs OTHERS)	233
Table 41 - Model fit (INP vs OTHERS).....	237
Table 42 - Results table for Bagging Models with balanced data.....	241
Table 43 - Examples used in the identification task.....	247
Table 44 - Summary of participants of the discrimination test	252
Table 45 - Summary of participants of the identification test.....	252
Table 46 - summary of the model simplification process (output of R's step() function).....	257
Table 47 - Output of the minimal adequate model, presenting the values of the binomial model's coefficients; the (CNT-EXP) level of the Pair factor was used for intercept. Uncertainty intervals (profile-likelihood) and p-values (two-tailed) computed using a Wald z-distribution approximation	259
Table 48 - Multinomial models - Complete model vs Model without triple factor interaction	262
Table 49 - Multinomial model's output - Identification task.....	263

LIST OF AUDIO FILES

Audio file 1 - afamcv01_174	44
Audio file 2 - afamcv01_174	44
Audio file 3 - afamcv01_174	45
Audio file 4 - afamcv01_174	46
Audio file 5 - afamcv01_174	46
Audio file 6 – afamd102_183-185 – COM.....	49
Audio file 7 - afamd101_111 – CMMs forming a compositional illocutionary pattern.....	53
Audio file 8 – afamd101_067 – TOP – Type 1	54
Audio file 9 - afamd101_080 – TOP – Type 2.....	55
Audio file 10 - afamcv04_138 – TOP – Type 3	56
Audio file 11 - afamcv01_025 – APC.....	58
Audio file 12 – afamd102_053 – APT.....	59
Audio file 13 – afammn05_010 – PAR	61
Audio file 14 - afamd103_106 – INT	63
Audio file 15 – afammn06_010 - Scanning Unit with retraction	64
Audio file 16 – apubmn01_285 – Empty prosodic unit	65
Audio file 17 – apubdl02_10 – Time-taking prosodic unit.....	65
Audio file 18 - afamcv02_174 - Unclassified unit	66
Audio file 19 – apubdl02_098 - Unclassified unit	66

Audio file 20 - bfamcv04_191-196.....	74
Audio file 21 - bpubcv03_123	75
Audio file 22 - bpubdl05_254.....	75
Audio file 23 - bfammn05_102	84
Audio file 24 – btelpv06_094	85
Audio file 25 – btelpv06_003	86
Audio file 26 – bfammn05_102.....	88
Audio file 27 - bfamdl01_201.....	89
Audio file 28 - bnatte03_093-094.....	96
Audio file 29 - bnatbu02_001-002	97

LIST OF EQUATIONS

Equation 1 – Z-scores	117
Equation 2 – Moving average filter	118
Equation 3 – Semitone cents between two frequencies	142
Equation 4 – Within-frame probability	142
Equation 5 – Semitone cents between two frequencies	143
Equation 6 – Between-frames probability	143
Equation 7 - Polynomial coefficients of the cubic function	177
Equation 8 - Precision	203
Equation 9 - Recall.....	204
Equation 10 - F1-score	204

LIST OF ABBREVIATIONS AND ACRONYMS

ALL	Allocutive
APC	Appendix of Comment
APT	Appendix of Topic
CMM	Multiple Comment
COB	Bound Comment
COM	Comment
COM_r	Reported Comment
CNT	Conative
DCT	Discourse Connector
EMP	Empty
EXP	Expressive
EVD	Evidentiator / Highlighter
f0	Fundamental frequency
i-COM	Interrupted Comment
INP	Incipit
INT	Locutive Introducer
LABLITA	Laboratorio Linguistico del Dipartimento di Italianistica dell'Università di Firenze
PAR	Parenthetical unit
BP	Brazilian Portuguese

PHA	Phatic
SCA	Scansion unit
TMT	Time-taking
TOP	Topic unit
UFMG	Universidade Federal de Minas Gerais
UPSaclay	Paris-Saclay University

LIST OF SYMBOLS

(*)	beginning of turn
(%)	beginning of a subordinated line
(ABC)	speaker code
//	terminal boundary
/	non-terminal boundary
(+)	interrupted sequence
(< >)	mark of overlapping speech
([/n°)	retraction
(&)	mark of interrupted word
(&he)	hesitation or filled pause
(" ")	citation
(hhh)	paralinguistic behavior
(xxx)	non-transcribed word
(yyyy)	non-transcribed audio chunk

SUMMARY

1	INTRODUCTION.....	35
1.1	Organization of this work	36
1.2	Summary of the research goals.....	37
2	THEORETICAL FRAMEWORK	39
2.1	Introduction.....	39
2.2	Why and how to segment the speech.....	39
2.3	The Terminated Unit: utterance or stanza.....	43
2.4	The pattern	43
2.5	The relationship between the prosodic and information patterning.....	47
2.6	The units of the Information Structure according to the L-AcT	48
2.6.1	The Comment (COM).....	48
2.6.2	The Topic (TOP).....	54
2.6.3	The Appendix of Comment (APC)	57
2.6.4	The Appendix of Topic (APT).....	59
2.6.5	The Parenthetic (PAR).....	60
2.6.6	The Locutive Introducer (INT).....	62
2.6.7	Scanning Units (SCA)	63
2.6.8	Empty, time-taking, interrupted, and unclassified prosodic units.....	64
2.7	Summary tables	66
3	REVISION AND DEEPENING OF THE PROPOSAL FOR DMs.....	69
3.1	Brief overview	69
3.2	Some features of DMs.....	70
3.2.1	Non-compositionality.....	71
3.2.2	Desemantization.....	73
3.2.3	Poly-functionality of DMs	76
3.2.4	Functions of DMs	77
3.2.5	Summary of the section.....	78
3.3	L-AcT's Discourse Markers framework history.....	78
3.4	The most recent proposal for DMs	80
3.4.1	The Incipit (INP)	81
3.4.2	The Conative (CNT).....	84
3.4.3	The Allocutive (ALL).....	86
3.4.4	The Expressive (EXP).....	87
3.4.5	The Highlighter (HGL/EVD).....	88
3.4.6	Summary table	89
4	MATERIALS AND METHODS.....	91
4.1	C-ORAL corpus.....	92
4.1.1	Core characteristics	92
4.1.2	Organization of the C-ORAL-BRASIL corpus.....	93
4.1.3	Segmentation	96
4.1.4	Text transcription	101
4.1.5	Morphosyntactic parsing	101

4.1.6	Minicorpus	101
4.1.7	Availability	102
4.2	DM Sampling	102
4.3	Data processing and annotation.....	108
4.3.1	Data preparation	108
4.3.2	Data annotation.....	108
4.4	Standardization of measures.....	112
4.4.1	Reference for the standardization of prosodic-acoustic parameters.....	112
4.4.2	Standardization of prosodic-acoustic parameters.....	113
4.5	Prosodic-acoustic parameters estimation.....	115
4.5.1	Features of intensity.....	116
4.5.2	Features of duration.....	116
4.5.3	Features of fundamental frequency (f0).....	118
4.5.4	Features of f0 variation	118
4.5.5	Alignment features	119
4.5.6	Features of f0 curves.....	120
5	EXTRACTING ROBUST F0 CURVES	122
5.1	Motivation.....	122
5.2	Pitch Detection Algorithms	123
5.3	Voicing decision model.....	126
5.3.1	Introduction.....	126
5.3.2	Methods.....	128
5.3.3	Models.....	133
5.3.4	Results.....	135
5.3.5	Conclusion.....	140
5.4	Dynamic programming algorithm.....	140
5.5	Output	146
5.5.1	Viterbi Algorithm.....	146
5.5.2	Voicing Decision Model.....	148
5.6	Algorithms used for the estimation f0 parameters	149
6	DESCRIPTIVE AND INFERENTIAL STATISTICS OF THE DISCOURSE MARKERS	150
6.1	Features of intensity.....	151
6.2	Features of duration.....	158
6.3	Features of fundamental frequency (f0).....	161
6.4	Features of f0 variation	165
6.5	Features of alignment.....	170
6.6	Features of fitted curves.....	176
7	CLASSIFICATION MODELS	185
7.1	Criteria for assessing different classification models	185
7.2	F0 curve fitting.....	186
7.3	Choosing a classification technique	198
7.4	Feature selection with leaps and bounds.....	208
7.4.1	Global model	209
7.4.2	ALL Against OTHERS.....	216

7.4.3	CNT against OTHERS	222
7.4.4	EVD against OTHERS.....	226
7.4.5	EXP against OTHERS	230
7.4.6	INP against OTHERS.....	234
7.4.7	Global model with Select K Best	237
7.5	Other models	238
8	PERCEPTUAL EXPERIMENTS	244
8.1	Introduction.....	244
8.2	Dataset of the discrimination task.....	245
8.3	Dataset of the identification task.....	246
8.4	Resynthesis	248
8.5	Participants	252
8.6	Discrimination and identification paradigms.....	253
8.6.1	Discrimination task	253
8.6.2	Identification task.....	254
8.7	Analysis of the discrimination task results	256
8.7.1	Binomial generalized model	257
8.7.2	Proportion of Match by Pair	260
8.8	Analysis of the identification task.....	261
8.8.1	Multinomial model	262
8.8.2	Effect of the presentation Modality	268
8.8.3	The lexeme	270
8.8.4	The functional CLASS attributed to the original stimuli	272
8.8.5	Interaction between presentation modality and lexeme	273
8.8.6	Interaction between modality and DM class.....	275
8.8.7	Interaction between lexeme and DM class.....	277
8.8.8	Triple interaction between MODALITY * LEXEME * CLASS.....	279
8.9	Discussion of results.....	280
9	CONCLUSION	283
10	REFERENCES	288
11	APPENDIX A - TRANSCRIPTION CRITERIA.....	302

1 INTRODUCTION

The main goal of this research is to study Discourse Markers (DMs) in spontaneous speech corpora. In the literature, the starting point of the study of DMs is usually the lexicon; one may choose, for instance, the lexeme *LIKE* and, based on syntactic, distributional, or contextual criteria, and study the functions that this item can assume. This procedure entails crucial problems to which I will return during this exposition. For now, it is enough to say that lexicon-based models have not allowed, to my knowledge, a framework that clearly defines, explains, and allows for the prediction of DMs' functions. The answer to this problem will be anchored on the prosody of Discourse Markers. To this aim, I will explain the importance of the prosodic parsing for speech and how discourse is organized by the theory underpinning this research, the Language into Act Theory (L-Act – Cavalcante, 2020; Cresti, 2000; Moneglia & Raso, 2014). We will discuss how DMs can be defined in a way that allows their prediction and how one can discriminate their functions based on their prosodic forms (Cresti, 2000; Frosali, 2008; Raso et al., 2022; Raso & Vieira, 2016). I will review previous works on the theoretical functions and forms of DMs and present a proposal containing five DM macro-functions. The foundations of a model aimed at explaining the mapping of prosodic forms onto the proposed DM theoretical functions will be outlined, along with exploring the extent to which this mapping is possible. Additionally, the validation of the proposal from statistical and experimental standpoints will be addressed.

A varied range of pitch, loudness, and speech rhythm acoustic correlates were estimated to model DMs. A frequent issue during data analysis concerns estimating and tracking the fundamental frequency (f_0), the correlate of voice pitch. This is especially true when dealing with spontaneous speech data. Recorded out of acoustically isolated and controlled settings, the speech signal can be rapidly degraded by the presence of reverberation, air-conditioners, fans, car engine noises, or other such phenomena. This may lead to Pitch Detection Algorithms (PDAs) inaccurately estimating f_0 values. Over the past 30 years, several methods and algorithms have been proposed to cope with such

problems. Each method displays different sensitivities to the diverse conditions degrading the speech signal. Another problem is that speech data recorded in uncontrolled settings may require researchers to manually revise f0 data and adjust various parameters provided by PDA algorithms. Since each recording setting exhibits different acoustic conditions, it can be assumed that each audio file will need different parametrization to achieve the most realistic f0 estimation and tracking. Ensuring the reproducibility of the work would require keeping track of all modifications made to f0 data and all parametrizations applied to each audio file analyzed. To streamline this potentially burdensome, time-consuming, and error-prone task, the proposed solution enhances f0 tracking by comparing the output of multiple PDA algorithms with a voicing decision model.

Another important aspect of this work concerns validating the proposed DM functions from a perceptual standpoint. For instance, the corpus-based approach presented by Lee et al. (2020), who extracted a large set of DMs from different corpora and classified them according to their functions, allowed the authors to observe correlations between the functions of DMs and their prosodic characteristics, across two languages (English and French) and several speech styles. Meanwhile, the perceptual validation of the importance of prosodic cues to the various functions of DMs are still rare in the literature: the work proposed in Didirková et al. (2019) being an important step for the validation of such relationships in French. In this research, a methodology for the validation of the relationship between DM functions and prosodic forms is proposed.

1.1 ORGANIZATION OF THIS WORK

Besides this introduction, this work is comprised of eight other chapters. The first two chapters are dedicated to the theoretical issues related to this research. The second chapter will review the theoretical foundations underpinning this research. I will show how one can segment the speech flow and how the L-AcT offers a framework accounting for its organization. In the third chapter, I discuss defining

and recognizing Discourse Markers in speech. I present the most recent proposal for the DM theoretical functions, their prosodic forms, and distributional constraints. Chapter 4 is dedicated to the methodological aspects of my work. In the fourth chapter, I present the spontaneous speech corpus from which DM instances were extracted. I also give further details on the methods I will use to extract and model the prosodic-acoustic parameters of the DMs. Chapter 5 is dedicated to the methodological endeavor of improving fundamental frequency estimation and tracking. I show how a model was trained to classify voiced/unvoiced regions and how f_0 estimations obtained from various PDAs produced an f_0 path. Chapters 6 and 7 are dedicated, respectively, to the presentation of the descriptive statistics and an exploratory data analysis of the DM instances found in the sample utilized in this work. Finally, in chapter 8, I outline the perceptual experiments designed to assess the degree of recognizability of DMs by means of their prosodic forms, and I present the results and issues of the experiments.

1.2 SUMMARY OF THE RESEARCH GOALS

The general objectives of this research can be summed up as it follows. The first objective is to present a new proposal for the L-Act's DM framework based on the prosodic form and to provide statistical and experimental evidence thereto. Another goal is to enhance f_0 estimations and tracking for audio files recorded in natural settings. The specific research goals are:

- (a) Review and deepen the classification of Discourse Markers and other short information units of L-Act's DM framework;
- (b) Implement a solution for the choice and tracking of f_0 candidates based on available PDAs and couple it with a voicing decision model;

- (c) Train and evaluate a supervised model to assess the degree to which a statistical model strictly based on prosodic information can classify DM observations into the proposed DM functions;
- (d) Run perceptual experiments aimed at evaluating the extent to which participants can discriminate and identify DM functions (by means of strictly prosodic manipulations).

2 THEORETICAL FRAMEWORK

2.1 INTRODUCTION

The theoretical framework underpinning this research is the Language into Act Theory (henceforth L-Act – Cavalcante, 2020; Cresti, 2000; Moneglia & Raso, 2014). The L-Act intends to be a corpus-driven theory, and its framework extends the Speech Act Theory (Austin, 1962). The theory's central tenets result from years of systematic observation and study of spontaneous speech corpora. These tenets are centered around the idea that to speak is to act in the world and that prosody plays a crucial role in conveying the functions of speech units. This chapter will present the theory's main principles and concepts. We begin with the importance of segmenting speech and establishing a reference unit of analysis. Then, we move on to some implications of speech segmentation. This is a precondition for defining what a Discourse Marker is. Its definition will be presented further ahead, but its ultimate goal is to allow the identification and prediction of the phenomenon. Firstly, we will present the reference units for analyzing the speech data according to the L-Act. Then, we show L-Act's proposal for the organization of speech – the Information Structure. Only after these steps we will be able to propose our definition of DM. Audio files used as examples of this and the ensuing chapters can be downloaded from [<SHARED MATERIALS THESIS>](#)¹

2.2 WHY AND HOW TO SEGMENT THE SPEECH

In this research, we are studying Discourse Markers based on data from spontaneous speech corpus. The first methodological question is how the speech can be segmented and what our primary reference unit is. Depending on the goal of any analysis, the speech may be segmented at different levels, in types of units of various sizes, each helping to

¹ https://1drv.ms/f/s!Ar5G4HnYDsd9goeGYdFY_6CL9ZID9hg?e=wWZKvQ

understand the relations at different linguistic levels (Izre'El et al., 2020) For instance, we can segment the speech into phones, syllables, words (whose definition is highly dependent on our approach – Blanche-Benveniste, 1997), or other higher-level linguistic entities (intonation units, utterances, turns, to name a few). The present research is inserted in a theoretical framework whose main objective is to explain how the same lexical content can be organized (or packaged) in different ways so that speakers achieve their communicative goals in spontaneous speech interactions (Raso & Cavalcante, 2022). Therefore, by reference unit, we mean the smallest communicative unit of speech. The explanations and exemplifications that follow try, thus, to highlight the importance of segmenting speech and establishing the basic units of reference from which our analyses will be carried out.

To show the importance of the subject, we can resort to one of the examples - and the discussion - presented in Izre'el et al. (2020). But before delving into it, we must introduce L-Act's central reference unit, in other words, the smaller communicative unit of speech. This unit is the *terminated unit* (TU):

Terminated unit

The minimal speech chunk that displays both pragmatic and prosodic autonomy.

The pragmatic autonomy means that the TU must convey at least one speech act (like an assertion, a calling, an invitation, an order, a question, or a warning, among many other possibilities). The prosodic autonomy means the speech chunk is perceived as complete, as concluded by a prosodic sign of terminality (a terminal boundary). This definition entails many consequences. For now, it is important to say that the TU is formed by at least one prosodic unit that conveys an illocution.

With this definition, we can analyze some linguistic relations within a string of words ripped off from its structure (syntactic or

semantic). Izre'El et al. (2020) propose the following sequence: *people give John the book I promised him*. This is not a natural example extracted from the corpora used for this research. Still, it helps, in the first moment, understand some implications of the prosodic segmentation of speech. We will present examples extracted from the corpus further ahead. Here, we will adapt the original rationale and assume only one kind of boundary, the one requested by the TU – the terminal boundary². We will mark this boundary with “//,” the same symbol adopted by the corpora used in this research³. We will assume the following arrangements of TUs (but others are possible):

Example 1

- (a) people give John the book I promised him //
- (b) people // give John the book I promised him //
- (c) people // give John the book // I promised him //
- (d) people give John the book // I promised him //

As can be observed, different segmentations result in different numbers of TUs – or, as in the case of (b) and (d), the same number but made up of different words. The segmentation is insufficient to impose an illocutionary value for each unit, but rather how many there are. The segmentation limits the potential illocutions we may have and, as such, constitutes a first step towards the interpretation of the sequences.

² Different proposals for the prosodic segmentation can be found in the literature. What is certain is that any kind of segmentation will imply the presence of a boundary, “either perceived or theoretically proposed and correlated to other kinds of phenomena” (Izre'El et al., 2020). As pointed out by Izre'El et al. (2020), two different perspectives can arise from the study of segmentation. The first one is focused on the units formed by boundaries – the prosodic units – and the functions they may carry. The second one is focused on the study of the acoustic cues that are associated with a boundary (Raso et al., 2020a, 2020b; Teixeira & Malvessi Mittmann, 2018).

³ The corpora of the C-ORAL family. We will give further detail on them in Chapter 4.1. C-ORAL corpus

Only after segmenting the speech and assigning an illocution to the TU, we can make morphosyntactic considerations. But before moving on to morphosyntax, let's say something about the illocutions that may be present in each TU. We can take, for the sake of simplification, the segmentation proposed for (a), (b), and (c) and suppose the following illocutionary values:

Example 2

- (a) people give John the book I promised him (Assertion) //
- (b) people (Calling) // give John the book I promised him (Order) //
- (c) people (Calling) // give John the book (Expression of surprise) // I promised him (Confirmation request) //

In (a), we have a single illocution, an *assertion*. In (b), we can have a *calling* followed by an *order*. In (c), we can have a calling followed by two confirmation requests, which could be paraphrased as "people! You really mean I should give John the book!? Did I promise him that?".

We can observe that establishing a TU will have many implications at the morphosyntactic level. One straightforward consequence is that morphosyntactic relations will have their primary domains of analysis within the prosodic unit, which in this case corresponds to the TU. Depending on how a sequence of words is segmented, the domain in which these words are related changes, and therefore, the relations change. For instance, in (a), *people* is the syntactic subject that gives *John* (the syntactic indirect object) *the book* (the syntactic direct object). The same is not true for (b) and (c). Here, the lexical item is used, on its own, to perform the illocutions of calling.

However, a TU may also display an internal organization. Words can be grouped into prosodic units that will, in principle, accomplish a communicative function. The prosodic units within a TU will be signaled by a non-terminal boundary. This kind of boundary is marked through a simple slash ("/"). We can take as an example the word sequence "in outer space research activities have been canceled". We will assume the following arrangements of prosodic units (but others

are possible):

Example 3

- (a) in outer space / research activities have been canceled //
- (b) in outer space research / activities have been cancelled //

As we can see, different boundary positions entail different local morphosyntactic relations and, therefore, different meanings. In (a) *research* specifies the subject *activities*, whereas in (b) it is adjunct specified by *outer space*. The segmentation, both with terminal and non-terminal boundaries, establishes the domain of relationship among words. Thus, speech cannot be analyzed unless it has been segmented first.

2.3 THE TERMINATED UNIT: UTTERANCE OR STANZA

So far, we have said that the TU must have at least one illocution and a terminal boundary. However, the TU may take on two forms: the *utterance*, when it is formed by a single pattern, or the *stanza*, when it is formed by more than one pattern (Cresti, 2010a). A *pattern* is made up of one illocutionary core unit and other optional units around it. When two or more patterns are juxtaposed (i.e., separated by non-terminal boundaries that convey continuation), we have a stanza. We will explore the concept of pattern a bit more in the next section.

2.4 THE PATTERN

A pattern may be of two kinds: *simple* or *compound*. A *simple pattern* is formed by a unique prosodic unit. This unit will necessarily be the one that carries the illocution. To illustrate the concept, we will show some examples extracted from the AE minicorpus (Cavalcante et al.,

2018)⁴. This minicorpus comprises texts sampled from the Santa Barbara Corpus of Spoken American English (SBCSAE – du Bois et al., 2000-2005) and was annotated in accordance with the methodological criteria adopted by the C-ORAL family corpora. The example below illustrates a simple pattern:

Audio file 1 - afamcv01_174

Simple pattern

KEN: [174] what kind of enzymes (Open question) //

LEN: [175] mainly digestive (Answer) //

In the example above, we have two TUs. Each one is formed by a simple pattern and performs one illocution. However, as previously discussed, the lexical content of a TU can be structured in more than one prosodic unit. The TU will, in this case, be performed through a *compound pattern*. The example below displays a TU whose pattern is formed by three prosodic units:

Audio file 2 - afamcv01_174

Compound pattern

FRE: [28] I put down on the card / you know / no cases (Assertion) //

⁴The examples shown in this chapter were extracted from the C-ORAL family corpora (see Chapter 4.1. C-ORAL corpus for further detail). They are identified by their ranking in the files attached to this work (Audio 1) and a code (afamcv01_174) that identifies their source file. The first letter in the code stands for the language of the minicorpus (a = American English, b = Brazilian Portuguese, f = French, l = Italian, p = European Portuguese, and s = European Spanish), the following three stands for the domain of interaction (fam = family/private, pub = public), and the final two letters represents the type of interaction (mn = monologue, dl = dialogue, cv = conversation). The number after the underscore indicates the rank of the TU inside the source text file. The transcription of each TU is introduced by a three-letter code identifying the speaker (KEN) and followed by the indication of the ranking.

Here, the compound pattern is formed by the prosodic unit carrying the illocution (an assertion) and by two other units. The sole mandatory unit is, thus, the last one. The two other units are optional. This means that the first two units can be disposed of without prejudice for the performance of the illocution. Moreover, if this TU were performed without one or both of the two non-terminal boundaries, its meaning would sensibly change. For instance, if there was no non-terminal boundary between *card* and *you know*, *you know* would have to be interpreted as a specifier of *card*. By hearing the corresponding audio, we clearly notice that this was not the speaker's intention. The speaker structured the pattern this way with a communicative goal in mind.

Thus, to sum up, the pattern is an assemblage of one or more prosodic units, one of which will necessarily carry the illocution (the illocutionary unit). Other optional non-illocutionary prosodic units can be added to the pattern. Optional units can occur both before, after, and, in some cases, even within the illocutionary unit, and the pattern can achieve a relatively complex structure. The non-illocutionary units will be functionally and subordinated to the illocutionary unit.

But often we can observe two or more patterns juxtaposed by a non-terminal boundary. They are, in such cases, *subpatterns* of the same TU. This TU is the stanza. Subpatterns can be assembled using the same constraints that apply to patterns. Each subpattern will thus have a core illocutionary unit, which can be complemented with other optional units.

Audio file 3 - afamcv01_174

Subpatterns

FRA: [176] I mean / I waited (Assertion) / and waited (Assertion) / and waited (Assertion) / and waited (Assertion) / and everyone had given up (Assertion)
//

In the stanza above, the pattern is formed by five subpatterns. Using

curly brackets, we can identify the beginning and end of each one:

Audio file 4 - afamcv01_174

Subpatterns

FRA: [176] { I mean / I waited (Assertion) / } { and waited (Assertion) / } { and waited (Assertion) / } { and waited (Assertion) / } { and everyone had given up (Assertion) // }

This stanza features a repetition of the same illocution with almost the same lexical content. We can easily observe how the illocutions are not performed in a sequence of TUs but rather in a unique TU whose patterns are linked by non-terminal boundaries. Only the first subpattern is complex. The illocution is performed by its second prosodic unit. The following four subpatterns are simple. The example below illustrates another stanza. Curly brackets once again delimit beginnings and ends of subpatterns, and illocutionary units are shown in bold:

Audio file 5 - afamcv01_174

Complex pattern

KIR: [81] { a potential for bringing over diseases / **is obviously there** / } { so / the thought was / okay / **let 's get some eggs** / } { Sea World San Diego / **has Gentus** / } { and / apparently / **just not enough eggs to share** // }

This stanza features four subpatterns. In this case, all of them are complex, i.e., formed by the illocutionary unit plus other optional units. In the next section, we will explain the relationship between the prosodic and the information patterning in accordance with the L-Act.

2.5 THE RELATIONSHIP BETWEEN THE PROSODIC AND INFORMATION PATTERNING

The L-AcT (Cavalcante, 2020; Cresti, 2000; Cresti & Moneglia, 2010; Moneglia & Raso, 2014) puts forth that the prosodic unit is the formal vehicle that conveys the function of the *information unit* (IU), i.e., a unit of the Information Structure (IS). The IS is a general term that includes concepts aiming to explain how the information is packaged (or organized) in the speech flow. L-AcT's approach to the IS assumes that speakers have at their disposal a limited inventory of IUs. The choice of an informational function is not constrained by the context (as in Krifka & Musan (2012)), but by what communicative value the speaker wants to give to a unit. Of course, the context has influence in the speaker's decision, but not in a deterministic way. This vision attributes to IS a strong linguistic status, since there are formal cues to recognize what the speaker wants to do, no matter the context. It is the speaker who decides, in the same context, what s/he wants to convey.

One of L-AcT's main principles is that there is a tendential *isomorphism* between the prosodic units and the IUs. Prosodic and information units are viewed as two dimensions of the same object. The L-AcT recognizes that prosody signals the boundaries of a pattern, segmenting it internally into interdependent units to which informational functions are associated. An IU will thus correspond in principle to each prosodic unit. The cases in which this tenet does not hold true will be explained at the end of this chapter. Besides, the specific informational function of each prosodic unit is marked by a specific prosodic form, as will be seen later.

Based on the observation of spontaneous speech corpora, a number of functions of IUs were identified and described. L-AcT's IS framework distinguishes two kinds of IUs: textual and dialogic units. Textual units are responsible for building up the text (the semantic and syntactic content) of the utterance. The dialogic units are, on the other hand, devoted to regulating the communicative exchange itself. The dialogic units correspond to what other frameworks call *Discourse Markers* (DMs). Unless otherwise specified, we will use DM and dialogic unit interchangeably. And since this is the very object of this research,

we will introduce L-Act's proposal for DMs in greater depth in a dedicated chapter. Here, I will focus on briefly presenting the textual IUs.

2.6 THE UNITS OF THE INFORMATION STRUCTURE ACCORDING TO THE L-Act

L-Act's IS framework identifies six main textual IUs: the Comment (COM), the Topic (TOP), the Appendix of Comment (APC), the Appendix of Topic (APT), the Parenthetical (PAR), and the Locutive Introducer (INT). The analysis of IUs is based on functional, prosodic and distributional criteria. This brings us to another important principle assumed by the theory: the form-function pairing. According to the L-Act, IUs have dedicated *prosodic forms*. A prosodic form is a set of prosodic parameters consistently associated with the conveyance of pragmatic functions of the same kind (Firenzuoli, 2003). Prosodic forms are typically described in terms of variations of fundamental frequency (f0), direction of f0 movement, f0 movement alignment, mean syllabic duration (articulation rate), and intensity. A prosodic form with which a prosodic unit is performed, guides, at the foreground, the conveyance of an information function. Besides, each IU will have distributional constraints or preferences with respect to the illocutionary unit. In the following subsections, we present the descriptions of each textual IU. At the end, a summary table containing tags, functions, and main references is also provided.

2.6.1 The Comment (COM)

The Comment (COM) is defined as the unit that conveys the illocution (Cresti, 2000, 2020; Raso & Rocha, 2016; B. Rocha, 2016; B. Rocha & Raso, 2016). This definition has three implications. The first one is that the COM is the sole necessary and sufficient unit for the performance of a TU. In other words, if a pattern is simple, the prosodic unit present will necessarily correspond to a COM. The second implication is that, in COM's case, the prosodic criterion leads to different prosodic forms.

Here, COM's prosodic form varies not in accordance with an information value *per se* but depending on the type of illocution carried by COM. Lastly, COM's distribution is free. Within a pattern, COM is the central reference unit with respect to which other IU's distributional constraints are described. The example below shows three examples of COM⁵ realized in dedicated prosodic units and filled with the same lexeme (love). Each COM (bold face) carries a different type of illocution conveyed through different prosodic realizations:

Audio file 6 – afamd102_183-187 – COM

DAR: [183] do what you want with the time you have // [184] learn / give / whatever //

PAM: [185] **love** //

%ill: directive (proposal)

DAR: [186] **love** //

%ill: expressive (doubt)

PAM: [187] **love** //

%ill: representative (conclusion)

COM's prosodic forms are described in terms of variations within a prosodic prominence called *functional nucleus*. This portion of the unit does not necessarily correspond to the whole syllabic extension of COM. The functional nucleus by itself is deemed to carry the prosodic characteristics responsible for signaling the illocutionary function. The functional nucleus is usually comprised of one or two syllables. The nucleus can be preceded by a preparation and/or followed by a coda, both of which do not carry information function but host the rest of the semantic content of the unit. Sometimes, the functional nucleus can be separated into two semi-nuclei; in this case, if necessary, a *linking portion* – also without information function – will lie between the two discontinuous parts of the same nucleus. More will be said

⁵ In the C-ORAL family corpora, IUs are annotated through a three-letter tag inside equal signs (=TAG=) that follows the boundary signs enclosing the referenced unit. The illocutionary value of COM is not annotated in the corpora, but it is here indicated following the annotation rules of the corpora (%ill: illocutionary value).

about COM and its forms when we talk about the illocution. The figure below shows f0 (blue line), intensity (red line) and the duration (the x-axis) of the three illocutions in Audio 4:

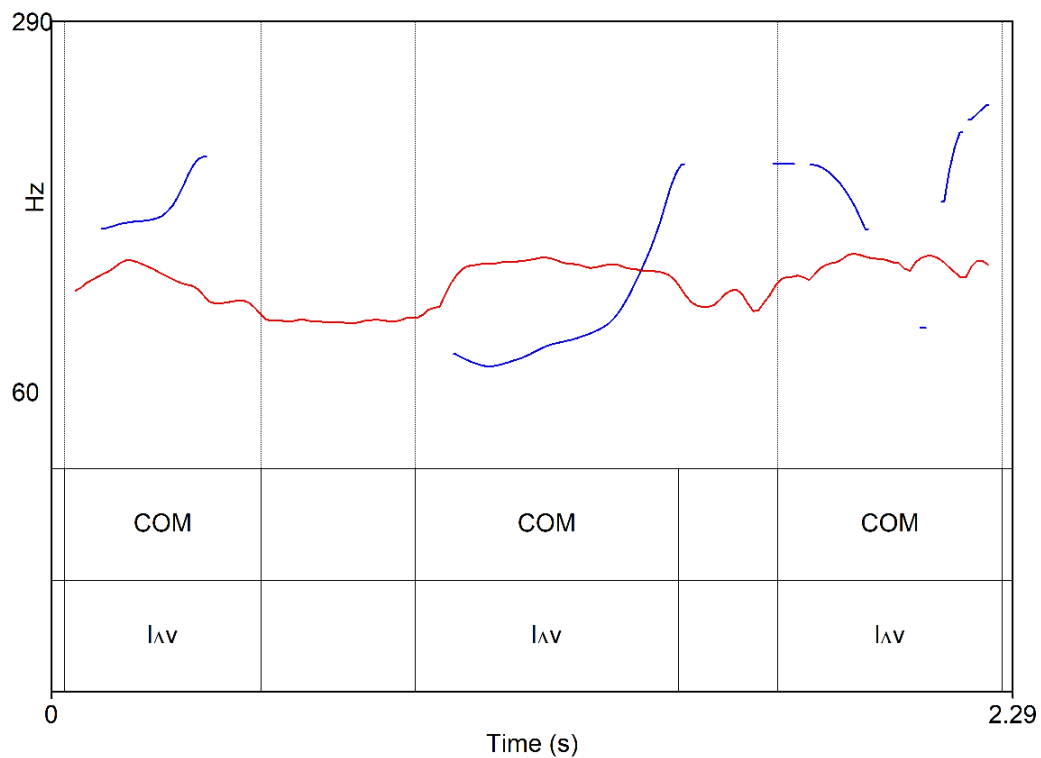


Figure 1 - Prosodic form of illocutions in Audio 4

We can see and hear that the three different illocutions have different prosodic forms, notwithstanding other aspects, discussed in the next section, also being important.

2.6.1.1 Pragmatic and cognitive parameters

COM is the unit responsible for carrying the prosodic form that will signal the illocution. The prosodic form will vary in accordance with what kind of illocution the speaker intends to perform. The

documentation of illocutions poses a strenuous challenge, with completion remaining an ongoing endeavor. A central question arises: which criteria should guide the classification of linguistic actions within the global context? The L-AcT advocates for the identification of pragmatic-cognitive parameters as descriptors of illocutions. It is imperative to recognize that an illocution cannot be exclusively defined by a particular prosodic form; rather, the form serves as a conveyance mechanism for a deeper conceptual content that still requires elucidation. The L-AcT proposes that the foundation for defining illocutions should rest upon the pragmatic-cognitive parameters – see, for instance, Moneglia (2011), as well as works by Raso & Rocha (2016), Rocha (2016a), and Rocha & Raso (2016). The pragmatic-cognitive perspective posits that certain illocutions possess inherent clarity, making formal differentiation unnecessary; examples include directives illocutions like *orders* and *instructions*. Conversely, for illocutions with closely aligned parameters, such as assertions and questions occurring in analogous contexts, the employment of formal prosodic distinctions becomes imperative.

The description of the pragmatic and cognitive parameters that allows the identification of an illocution plays a central role in the methodology utilized within L-AcT's research groups. It is important to mention that such descriptions must take into account the smallest possible number of pragmatic and cognitive parameters so as to avoid unnecessary overspecification (Rocha, 2016b). Furthermore, the choice of parameters and respective specifications must be carried out experimentally (see Rocha, 2016; Raso & Rocha, 2016a; Raso and Rocha, 2016b, for an empiric methodology on the identification of illocutions). A number of at least five parameters have been observed to be relevant to the distinction of illocutions cross-linguistically (Moneglia, 2011; Rocha, 2016; Raso & Rocha, 2016; Cresti & Fujimura, 2018; Cresti & Moneglia, 2018). They are listed in Table 1:

Table 1 - Pragmatic and cognitive parameters

Type of Parameter	Parameter
Communication	Channel
	Attentional horizon
	Focus
	Context
	Reference object
Proxemics	Space relations between participants and their movements
	Gesticulation
	Gaze
Social	Speaker's roles and conditions
	Addressee's roles and conditions
Speaker activity	Intentional values
	Speaker's commitment to the truth
	Speaker's affective involvement
Expected effects	Conventionally expected effects on the addressee
	Conventionally expected effects in the context
	Fulfillment time
	Benefit

(Adapted from Cresti, 2020)

For instance, many illocutions cannot be performed when the communication channel is not open, such as in the case of questions, confirmation request, and presentations. The need to perform such illocutions when the communication channel is closed may elicit the performance of a *patterned illocution* where the first illocution will be a *call* aimed at opening the channel. I talk about the patterned illocutions in the following subsection.

2.6.1.2 *Patterned illocutions*

I said before that each pattern contains one COM together with other optional IUs. However, the pattern may sometimes present more than one illocutionary unit. When this is the case, the illocutionary units are

called Multiple Comments (CMM). CMMs form a chain of two or more illocutionary units targeting a unified rhetorical effect. They thus seem to result from a single planification by the speaker. Chains of CMM are signaled by strongly conventionalized prosodic patterns, constituting a *patterned illocution conventionalized in order to achieve a holistic rhetorical effect* (Panunzi & Gregori, 2012; Panunzi & Mittmann, 2014). Although each CMM performs its own illocution, the illocutionary pattern must be interpreted as a whole. Chains of CMM form one unique nuclear pattern around which other optional non-illocutionary IUs can be added. Typical compositional illocutionary patterns are lists, comparisons, and requests of confirmation.

Audio file 7 - afamd101_111 – CMMs forming a compositional illocutionary pattern

BER: do I get it /=CMM= or not //CMM=
%ill: Request of confirmation

2.6.1.3 Stanzas

As aforementioned, a stanza (Cresti, 2010b) is a type of TU formed by patterns juxtaposed by non-terminal boundaries. In a stanza, the illocutionary unit of each pattern enclosed by non-terminal boundaries is annotated as a Bound Comment (COB). The illocutionary unit of the last pattern (the one enclosed by a terminal boundary) is annotated as a regular COM. Instead of having a unified rhetorical effect, COBs – and stanzas – are rather the product of the speaker's flow of thought. They tend to occur longer turns, when the actional activation is lower, like in monologic speech. In these situations, it is the semantic content that takes on the central role, leading to a weakened sequence of illocutions typically of the same class, such as in the example below, in which the illocutionary units pertain to the assertive class:

Audio 6 – afammn03_124 – COBs forming a stanza

ALA: [124] so I 'm driving up to the house /=COB= and there 's a car in front

of me /=COB= and the guy is just like sitting there /=COB= in the middle of
the road /=COB= and he 's not moving /=COB= and you know / I wanna park
the car // =COM=

2.6.2 The Topic (TOP)

Functionally, the Topic (TOP) provides a domain of identification (individual, spatial, temporal, etc.) for the interpretation of the illocution conveyed by COM. If not preceded by a TOP, COM must be interpreted in accordance with a domain given in the context. TOP allows for the detachment from the context (Hockett, 1958). Prosodically, TOP is also characterized by the presence of a *functional nucleus*, which carries the function. TOP's functional nucleus can take on three forms (Cavalcante, 2020; Firenzuoli & Signorini, 2003; Raso et al., 2016; Raso & Cavalcante, 2021; Cavalcante, Raso, Barbosa, to appear). Type 1 is characterized by a rising-falling fundamental frequency (f0) movement in the last stressed and post-stressed syllables. The following examples were adapted from Cavalcante (2020):

Audio file 8 – afamd101_067 – TOP – Type 1

[67] once I get my experience /=TOP= I'll be up there too / in the top-four
salesmen

//

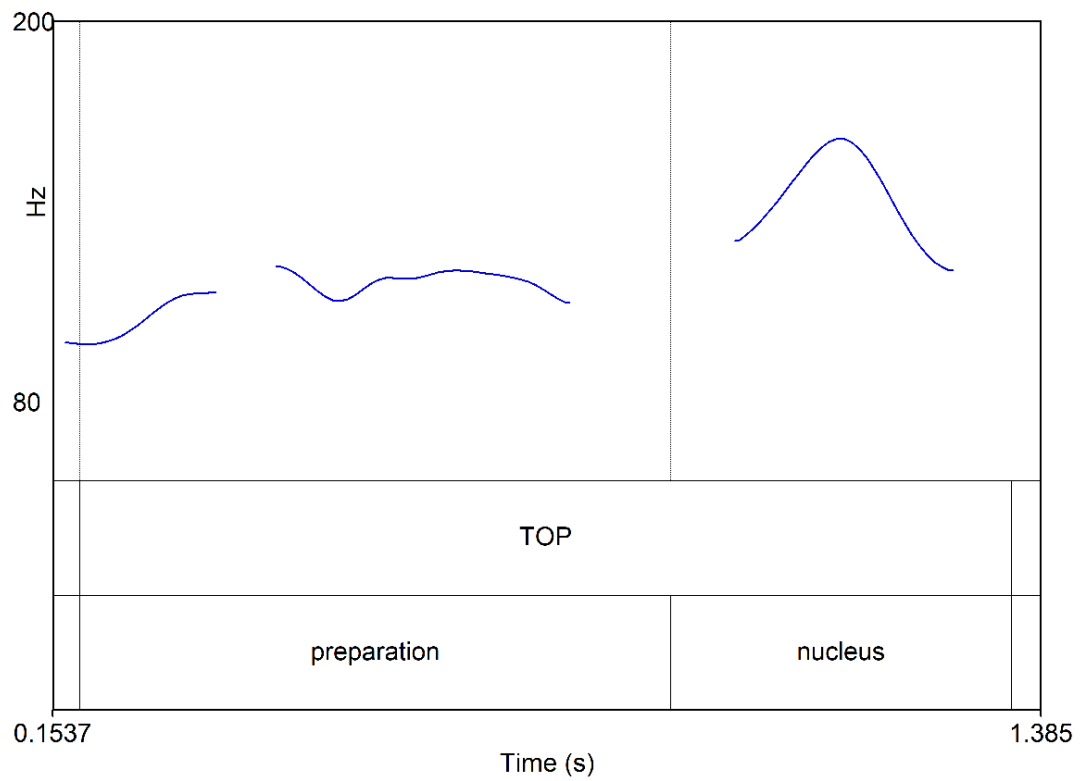


Figure 2 - Type 1 TOP

Type 2 is marked only by a rising fundamental frequency (f0) movement in the last stressed and post-stressed syllables.

Audio file 9 - afamd101_080 – TOP – Type 2

XXX: [80] but in a sense /=TOP= I need a [/1] some type of steady income //

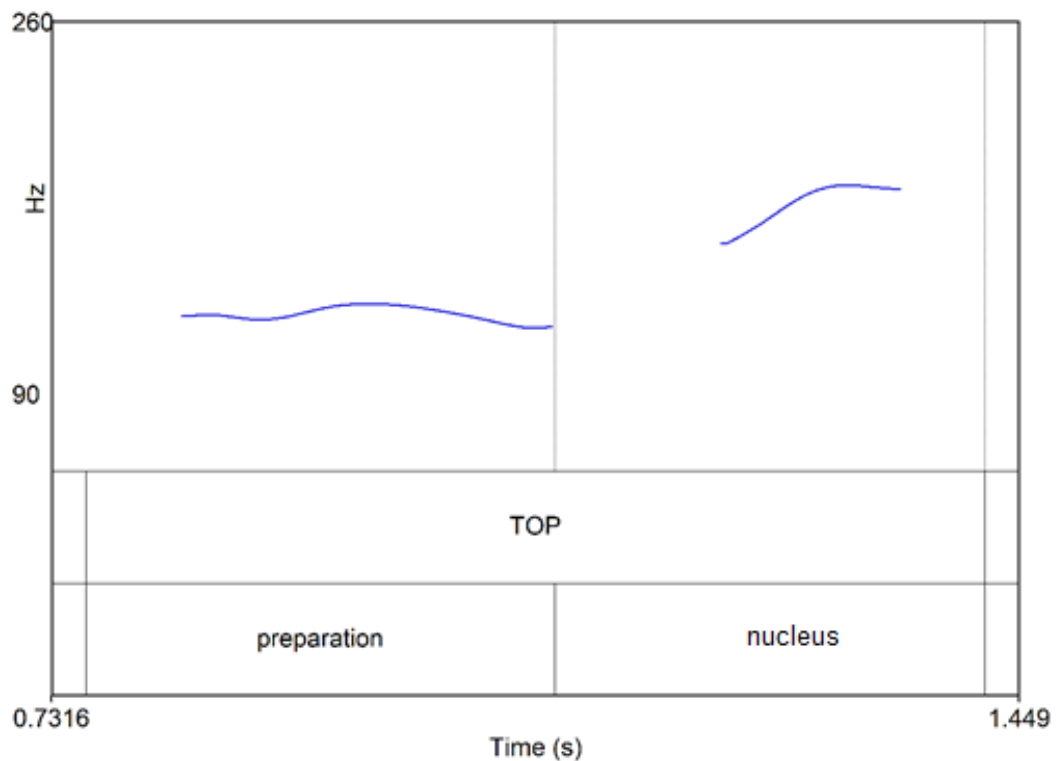


Figure 3 - Type 2 TOP

Type 3 has two semi-nuclei – often discontinuous: the first semi-nucleus displays high to extra-high f_0 values, and the second semi-nucleus has lower f_0 values. When they are discontinuous, the two semi-nuclei are separated by functionally inactive syllables called *linking portions*. The linking portion corresponds to what to the preparation of the other two types of TOP.

Audio file 10 - afamcv04_138 – TOP – Type 3

XXX: [138] when Mary tells me to get a sleep over the weekend /=TOP= you know I need to get sleep over the weekend //

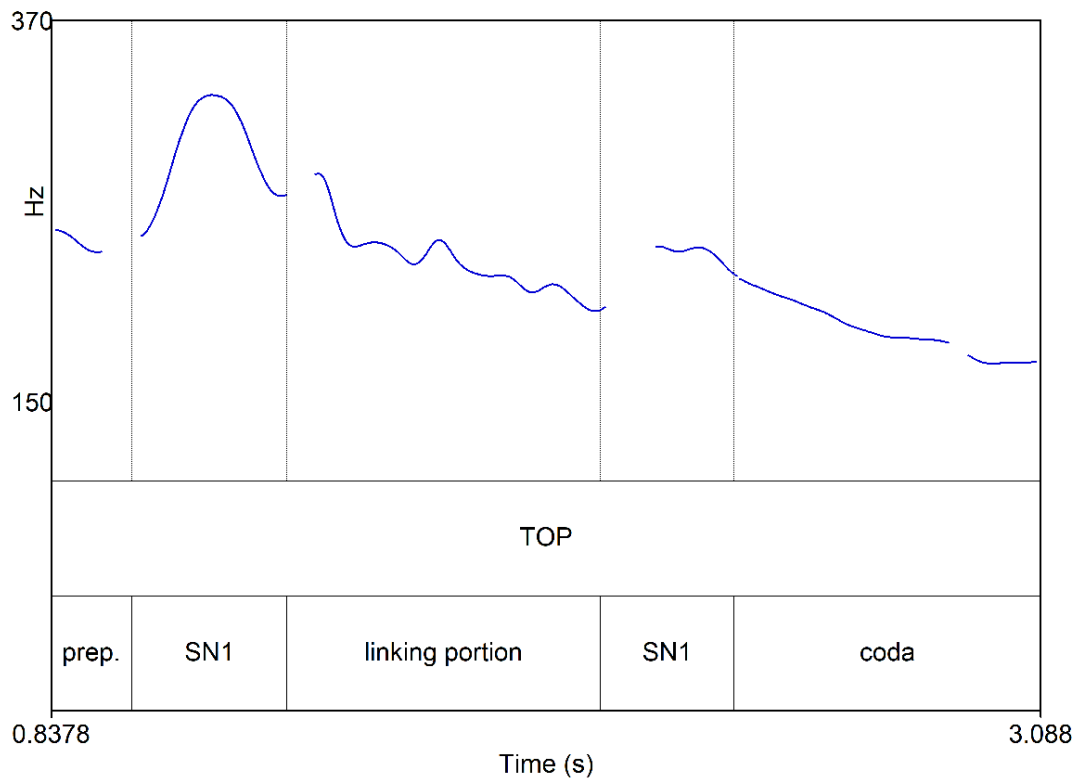


Figure 4 - Type 3 TOP

All three types feature syllable lengthening and higher intensity in the nuclei. Types 1 and 2 display the most prominent lengthening. Distributionally, TOP always occurs before COM.

2.6.3 The Appendix of Comment (APC)

Functionally, the Appendix of Comment (APC) integrates COM with textual content that usually corresponds to information already available in the context. Although integrating COM, the APC does not contribute to the performance of the illocution. If APC is cut off, the pragmatic autonomy of COM remains unchanged. Prosodically, the

APC is characterized by a flat or falling f0 movement without a functional nucleus (Cavalcante, 2020; Moneglia & Raso, 2014). Distributionally, it occurs always after COM. The example below brings an example of APC in utterance 25. We provide some context before to make it possible to see how the referent *all places* is already given in stanza 23⁶:

Audio file 11 - afamcv01_025 – APC

KEN: [23] but the whole town /=TOP= still has the old Mexican plaza /=COB= and the Mexican governor / general's house /=TOP= was right there /=COB= and / <and the church /=TOP= and that kind of thing is> / you know / right in the center of Sonoma // =COM=

JOA: [24] and that 's like the main street /=COM= you know //

JOA: [25] Sonoma /=COM= of all places // =APC=

⁶ Overlapping speech is transcribed within angle brackets (<speech>). IUs that were not introduced yet do not receive annotation in this example.

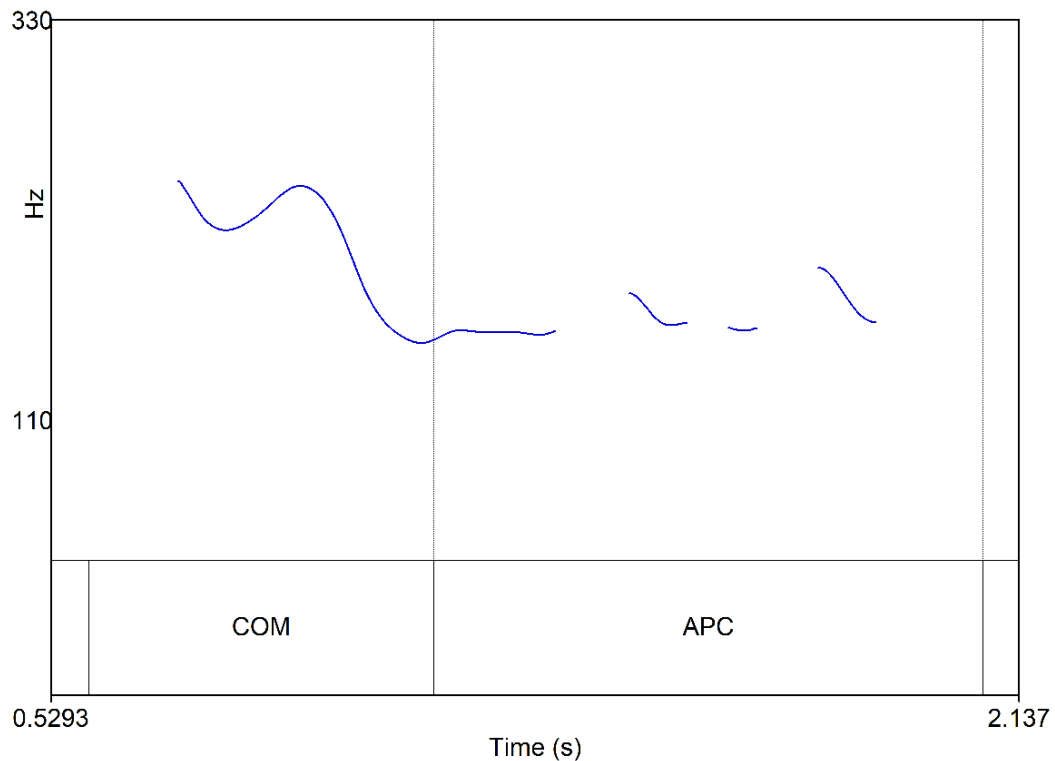


Figure 5 - APC

2.6.4 The Appendix of Topic (APT)

Analogue to the APC, the Appendix of Topic (APT) is functionally characterized by providing TOP with textual content integration. Differently from APC, this integration is not contextually given information; it rather supplements the domain of application of the illocutionary force identified by TOP. Its prosodic form seems, on the other hand, to be a bit more complex than that of the APC. Sometimes, it reproduces the f_0 contours of TOP in a smaller range and without a functional nucleus; sometimes, APT has a falling f_0 movement (Cavalcante, 2020). Distributionally, APT always occur after the TOP.

Audio file 12 – afamd102_053 – APT

PAM: [53] the things I know most /=TOP= **about life and death** /=APT= come from

[/1]=SCA=⁷ from /=SCA= my grandmother //COM=

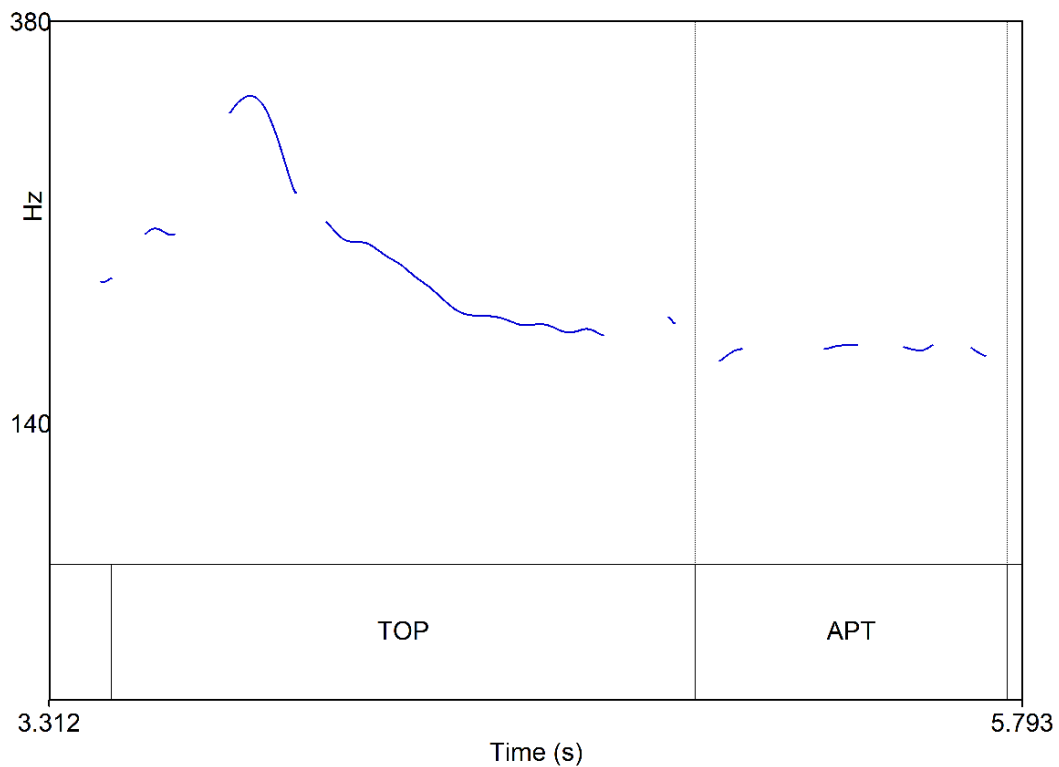


Figure 6 - APT

2.6.5 The Parenthetic (PAR)

The Parenthetic (PAR) has a metalinguistic function. It delivers a commentary on the content of its hosting pattern (Tucci, 2004a, 2009). PARs are frequently used as a modalizing mechanisms, expressing the speaker's point of view on the content of the pattern. Prosodically, PAR

⁷ Unit to be introduced in 2.6.7.

is characterized by an overall flat and low f0 profile, typically higher articulation rate, and low intensity with respect to the neighboring IUs (Tucci, 2004b). It is frequently followed or preceded by silent pauses. Distributionally, it can occur in any position – even within another textual IU – except for the beginning of the pattern. The example below illustrates a PAR that is embedded in the first COB⁸:

Audio file 13 – afammn05_010 – PAR

COR: [10] <and then like> [/3] =EMP= and then they 'll like /=INT= take these /=SCA= butt plugs /=COB= **or whatever you wanna call 'em /=PAR=** and they 'll shove it up their anus /=COB= and /=AUX= they have to walk around with it //COM=

⁸ IUs interrupted by another intervening IU and then resumed are signaled with the prefix “i-” before the tag of its interrupted chunk.

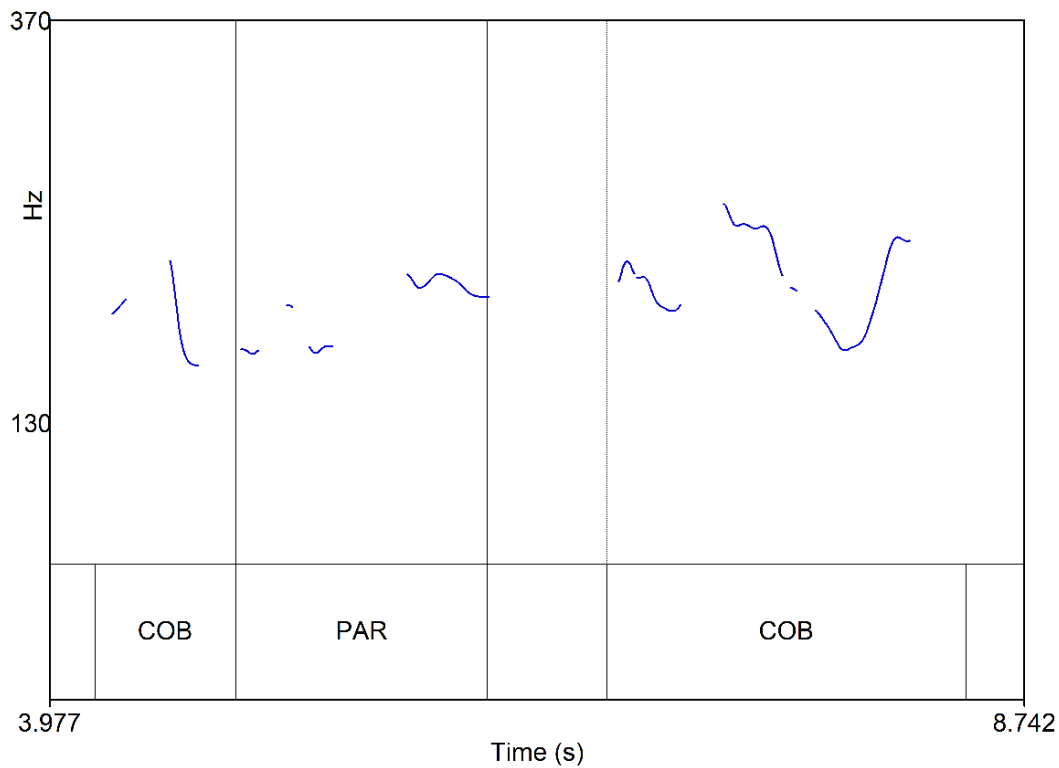


Figure 7 - PAR

2.6.6 The Locutive Introducer (INT)

Functionally, the Locutive Introducer signals that the illocution that follows has pragmatic coordinates (individual, temporal, and spatial) distinct from those of the unfolding TU. Most frequently, it is used to introduce meta-illocutions such as reported speech, but it can also signal illocutions containing spoken thoughts, lists, emblematic exemplifications, to name a few possibilities. It can also introduce lists of PARs. Prosodically, it is characterized by a falling f_0 profile at the end of the unit, a higher articulation rate and pronounced phonetic reduction (Maia Rocha & Raso, 2011; Maia Rocha, 2011; Toledo, 2024). Besides, it tends to feature a sharp prosodic contrast with respect to the introduced IU. Distributionally, it always occurs before the introduced IU. The example below illustrates INT introducing a

reported speech⁹:

Audio file 14 - afamd103_106 – INT

ANE: 106] and I ate the other one /=COB= then half of the other one /=COB=
it was like /=INT= whoa //COM=

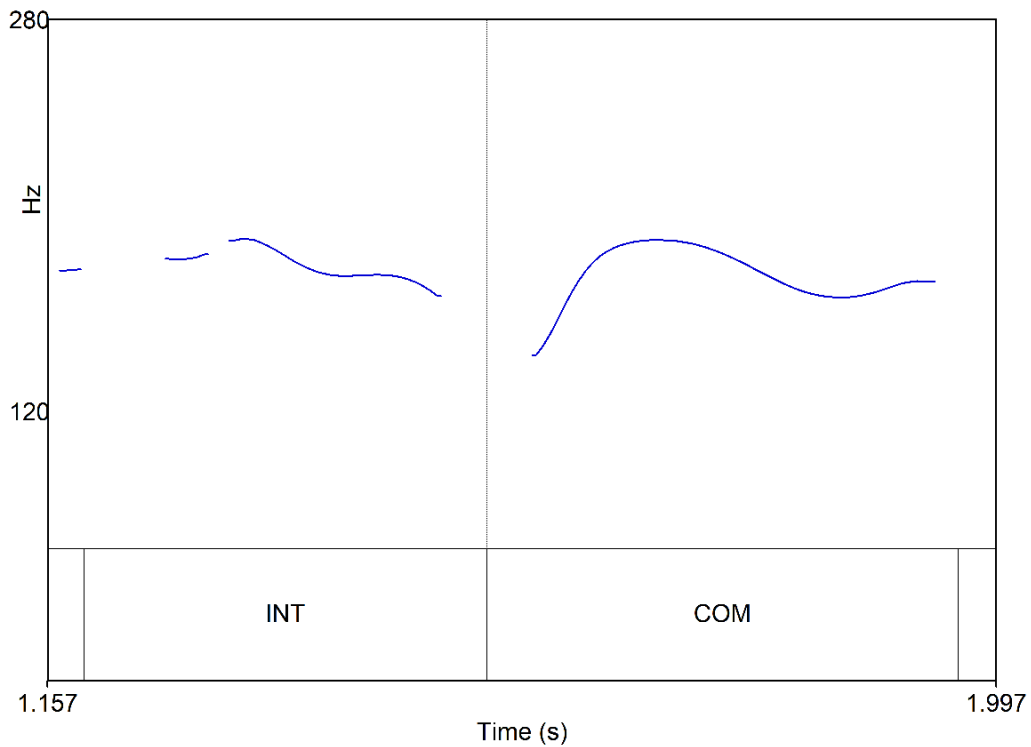


Figure 8 - INT

2.6.7 Scanning Units (SCA)

We said that there is a tendential isomorphism between the prosodic

⁹ IUs part of reported speech are annotated through the suffix “_r” after its tag.

unit and the IU. This principle is flouted in some cases. Sometimes, an IU can be realized by means of more than one prosodic unit. This may be caused by different reasons. Most frequently, this is due to dysfluencies or caused by articulatory reasons (the content of an IU may be too big to be articulated in one single prosodic unit). Less often, splitting the IU into multiple prosodic units is made for rhetorical purposes. In such cases, the prosodic units coming before the last prosodic unit of the IU receives the annotation of Scanning Units (SCA). The prosodic form that carries the information function will be always at the last prosodic unit of the IU.

In some situations, the speaker may replan her speech program, potentially leading to an SCA and the *retraction* of some words. A retraction occurs when the speaker makes minor adjustments to the initial program by withdrawing some words.

Audio file 15 – afammn06_010 - Scanning Unit with retraction

JIM: [10] and /=AUX= the way it 's marketed /=COB= and the way we 're [/?]=SCA= we develop needs for it //COM=

In the example above, the content *we're* is withdrawn in favor of *we develop*. The sign [/?] indicates both the occurrence of a non-terminal boundary and the retraction of two words before. Notice that the IU here is not abandoned but rather split into two prosodic units; a portion of the SCA unit – *and the way* – is not retracted. The IU is realized through two prosodic units: SCA and COM.

2.6.8 Empty, time-taking, interrupted, and unclassified prosodic units

Sometimes, the speaker retracts the content of an entire prosodic unit. This may be caused by the need to restart the speech planning. The fully retracted prosodic unit is annotated as an *empty prosodic unit* (EMP).

Audio file 16 – apubmn01_285 – Empty prosodic unit

AMY: [285] are they [/2]=EMP= I mean /=AUX= are they being hesitant about it //COM=

This case is not to be confounded with a *meaningful repetition*, which can occur in speech (Cavalcante, 2020). In meaningful repetitions, the information function is identifiable through a clearly realized prosodic form and its rhetorical effect. In the example above, the empty prosodic unit seems driven by a *speech dysfluency*. Speech dysfluencies are non-pathological hesitations that lead to phenomena such as repetitions, repairs, and filled pauses (time-taking vocalizations). Such dysfluencies are frequent even in fluent speech, and can create prosodic units without an information function.

The sound used in filled pauses (time-taking vocalization) may vary across languages. The corpora of the C-ORAL family signal these vocalizations with the generic sign *&he*. When it specifically carries one such vocalization, the prosodic unit receives the annotation TMT (Time-taking). The example below illustrates one such filled pause:

Audio file 17 – apubdl02_10 – Time-taking prosodic unit

LAR: [10] &he /=TMT= I do n't know //COM=

Sometimes, speakers may also stop their speech program by beginning a new TU or by abandoning the turn – for instance, when they are interrupted by another participant. When this is the case, the boundary created is annotated with a + sign. And, if, for any reason, the information unit cannot be recognized, the IU receives the tag Unclassified IU (UNC). This may be the case when the IU is interrupted, such as in Audio file 13, or when the speech chunk is perceived as an IU, but its content and communicative value cannot be recognized due

to overlapping, background noises, and muffled or whispered speech. Audio file 14 illustrates the latter case. Here, the xxx sign marks that one word was not recognized during the corpus transcription task¹⁰.

Audio file 18 - afamcv02_174 - Unclassified unit

BET: [174] no /=COB= I +=UNC=

Audio file 19 – apubdl02_098 - Unclassified unit

LAR: [98] <xxx> +=UNC

2.7 SUMMARY TABLES

Table 2 below presents a summary of the textual IUs proposed by the L-Act, as well as their tags, functions, and main references:

Table 2 - Synthetic table of the textual IUs assumed by the L-Act, their functions and main references

IU	TAG	FUNCTION	REFERENCES
Comment	COM	Conveys the illocution. It is the necessary and sufficient unit for the realization of the terminated unit	(Cresti, 2000; 2020; Moneglia & Raso, 2014; Raso & Rocha, 2016a; 2016b; Rocha, 2016)

¹⁰ The other signs used in the transcription of the C-ORAL family corpora will be dealt with when we present the corpora used for this work.

IU	TAG	FUNCTION	REFERENCES
Multiple Comment	CMM	Patterned illocutions that produce a conventionalized rhetoric effect; it is built upon a single illocutionary pattern	(Panunzi & Gregori, 2012)
Topic	TOP	Identifies the domain of identification – spatial, temporal, individual – for the interpretation of the illocutionary force carried by COM	(Firenzuoli & Signorini, 2003; Signorini, 2005; Mittmann, 2012; Rocha, 2012; Cavalcante, 2016; Raso et al., 2017; Cavalcante, 2020; Raso & Cavalcante, 2021)
Bound Comment	COB	The illocutionary unit of a sequence of subpatterns juxtaposed by non-terminal boundaries. It forms a stanza.	(Cresti, 2010)
Appendix of Comment	APC	Adds textual content to a COM unit, often corresponding to given information	(Moneglia & Raso, 2014; Cavalcante, 2020)
Appendix of Topic	APT	Adds textual content to a TOP unit	(Moneglia & Raso, 2014; Cavalcante, 2020)
Parenthetic	PAR	Delivers a metalinguistic commentary on the content of the TU.	(Tucci, 2004, 2009)
Locutive Introducer	INT	Signals that the following IU must be interpreted according to pragmatic coordinates other than those of the unfolding terminated unit	(Maia Rocha, 2010; Maia Rocha & Raso, 2011; Toledo, to appear)

Table 3, on its turn, presents a summary of tags and uses given to other

prosodic units that either do not bear an informational value or that could have their value identified:

Table 3 - Synthetic table of tags given to other prosodic units

UNIT	TAG	USE
Scanned Unit	SCA	Identifies cases in which the IU is realized through more than one prosodic unit. The first prosodic units of the IU will be annotated through SCA. Only the last prosodic unit will receive the tag of the corresponding IU.
Empty Unit	EMP	Identifies units that the speaker intends to withdraw. Mainly (but not only) used for repairs and repetitions that do not purport a rhetorical effect.
Time-taking Unit	TMT	Identifies prosodic units created by filled pauses (time-taking vocalizations)
Unclassified Units	UNC	Identifies prosodic units that for some reason (overlapping speech, background noise, muffled or whispered voice) could not have their informational value recognized.

In the next chapter, I will focus on deepening the L-Act proposal for Dialogical Units. I will explain why this proposal accounts for Discourse Markers better than other approaches that depart from the lexicon, context and/or syntactic structure.

3 REVISION AND DEEPENING OF THE PROPOSAL FOR DMs

The main goal of this chapter is to propose a revised framework for interactional Discourse Markers (DM). A substantial body of literature about DMs has been produced since the 1980s. However, it is still necessary to define what a DM is. More specifically, we lack a satisfactory response to two important questions. First, how can one determine whether a small expression or lexical item function as a DM? Secondly, how can one determine a lexical item's precise function once one has determined that it behaves like a DM? One major issue that biases the studies on DMs is that the lexicon is typically the starting point, with rare and partial exceptions. In the attempt to respond to the first question, it is contended that the prosodic cues, not the lexical filler, are the means by which DMs can be identified. Next, I will tackle the second question and demonstrate how corpus analysis enables us to determine five prosodic forms for five distinct DM functions. Before doing that, I will (a) briefly review the literature on DMs and (b) comment some defining features as per the literature.

3.1 BRIEF OVERVIEW

Discourse Markers have been on the agenda of various branches of linguistic studies and allied fields for as long as since the early 1980s. One important milestone on their study is Schiffrin (1987), but it is not until the mid-1990s that Discourse Markers began to come to its own (Brinton, 2010) with research focusing on English (Jucker, 1997; Traugott, 1995) but also on other languages such as Catalan (Cuenca & Marín, 2012), Chinese (Biq, 1990), Croatian (Dedaić, 2005), Danish (Emmertsen & Heinemann, 2010), Dutch (Mazeland & Huiskes, 2001), Estonian (Keevallik & Vint, 2012), Finnish (Hakulinen, 1998), French (Cadiot et al., 1985), German (Abraham, 1991), Hebrew (Maschler, 1997), Hungarian (Dér & Markó, 2010), Icelandic (Hilmisdóttir, 2011), Indonesian (Rofiq, 2018), Italian (Bazzanella, 1990), and Latin (Kroon, 1997). These bulk of research has explored the cognitive, expressive, social, and textual aspects of

Discourse Markers. With a few exceptions, the literature identifies the scope of the research departing from lexical items or small expressions presenting some syntactic characteristics. DMs are seldom defined, and, not rarely, the category encompasses a large number of other phenomena that, in our view, could be better accounted for as yet other phenomena. I will come back to this point further ahead.

Discourse Markers have been studied from different perspectives by different groups of scholars. The definition of DMs seems to be related to research interests and theoretical frameworks supporting the study. Schiffrin's (1987) initial work defined DM as "sequentially dependent elements which bracket units of talk" and proposed that they can be considered as a set of *linguistic expressions* comprising members of word classes as varied as *conjunctions, interjections, and adverbs*. Fraser (1999) defines DMs as a class of lexical expressions drawn primarily from the syntactic classes of conjunctions, adverbs, and prepositional phrases. Maschler's approach views all DMs as metalinguistic units, emphasizing this characteristic as their basic defining feature. Heine et al. (2019) view DMs as invariable expressions that are syntactically independent from their environment, typically set off prosodically from the rest of the utterance, and their function is to relate an utterance to discourse situation.

We can see that what has been studied as Discourse Markers can comprise a varied number of phenomena. DMs can have interactional, metalinguistic/metacomment, and textual cohesive macro-functions. It can even work as frame-shifting device. In the next section, I comment on some of the formal features and characteristics of DMs typically found in the literature.

3.2 SOME FEATURES OF DMs

The literature points to some formal features and characteristics that help define, identify and predict Discourse Markers (DMs). Some characteristics pointed out by the literature, according to Raso et al. (2022) and Raso & Ferrari (2020) are:

- (a) DMs are lexical items or small expressions that do not combine on the semantic and syntactic level with the rest of the utterance. This means that they are not properly a part of the propositional content and are, therefore, non-compositional items;
- (b) The lexical items or small expressions functioning as DMs lose (at least partially) their semantic meaning and acquire a pragmatic function;
- (c) DMs are polyfunctional; this statement may be used in two different senses: in the first one, it means that one DM occurrence may have one or more functions at the same time; in the second one, it means that a lexical item or small expression may take on different functions in different occurrences depending on the context;
- (d) A varied range of DM functions is found in the literature. By way of example, we can mention functions concerned with linguistic modality, illocution, conative function, turn-taking devices, and politeness, metalinguistic, and that is not an exhaustive list.

I would like to address these characteristics so as to evaluate to what extent they are adherent to L-Act's framework. Arguments are presented in the order in which these characteristics are set out above.

3.2.1 Non-compositionality

First of all, we agree with (a) that DMs are non-compositional items

both from semantic and syntactic standpoints. They do not combine with the propositional content of the utterance. The first question we should be concerned by is, thus, (a) how the non-compositionality is signaled in speech and (b) what features allows the addressees to understand the differences between the pairs of realizations of the three examples that follow:

Compositionality – Example 1¹¹

(a) God save the queen!

and

(b) God, save the queen! (Where God is an exclamation and the rest of the utterance performs an order)

In the first case, GOD is the subject of the sentence, and it is in a relation of compositionality with the text of the utterance. In the second example, GOD could be replaced by very different lexical items used as, for instance, exclamations or imprecations (Raso et al., 2022). In any case, we need formal criteria of non-syntactic nature to ascertain which interpretation to follow since, here, there is nothing neither in lexicon nor in syntax that say that we may have a boundary between GOD and SAVE.

Compositionality – Example 2

(a) Tipo meu deus.

Like my god.

(b) Tipo, meu deus!

¹¹ Example adapted from Raso & Ferrari (2020).

Like, oh my god!

TIPO (type) is canonically a noun in Portuguese and not a comparative connective such as LIKE. In example (a), it is used as a grammatical item: *como meu deus* (like my god), a non-canonical use that is frequently observed in spontaneous speech. Here, the item is syntactically and semantically compositional with rest of the utterance. In example (b), TIPO can be used with an interactional function. The same logic can be applied to small expressions like *I mean*:

Compositionality – Example 3

- (a) I mean I'm not going.
- (b) I mean, I'm not going.

In (a), a compositional use may lead to the paraphrase “by what I said I meant that I'm not going”, while in (b) the speaker may just want to draw the interlocutor's attention.

These examples were brought up to show that the lexical item (or the small expression) and the syntactic structure are far from enough to mark the loss of compositionality. To say with certainty that the compositionality is broken, we must take into account, firstly, the prosodic segmentation. In the data analyzed in this research, the interruption of compositionality is marked by a prosodic non-terminal boundary. We can now move to the second item of the list of DM properties.

3.2.2 Desemantization

Discourse Markers, in the sense that interests this research, are lexical items or small expressions that lose their semantic value. I want to show how an analysis that considers the prosodic form can help us

evaluate the semantic emptiness. The first condition is that the lexeme or small expression be isolated in a prosodic unit (i.e., that it be set off from the rest of utterance by a non-terminal boundary). This gives a first clue that the unit may be non-compositional. This discussion may be a bit longer, but it helps us respond to the last two points of our list – especially when I talk about DM functions.

For this discussion, it may be useful to take some examples of the same lexical item occurring in different contexts, either in prosodic isolation or not. I will resort to an item that may occur in many different informational contexts: ASSIM (like this). This item can be used with its full semantic value but also with interactional functions. This causes the lexeme to occur in units of widely varying informational values. Such a feature is desirable if we aim to show how the same lexeme can take on different functions depending on its prosodic realization. For the sake of space, I will not illustrate all the different functions ASSIM can assume but rather three different contexts it can occur in. A more detailed list of functions ASSIM can take on and the pragmatic implications can be found in Raso & Santos (2020). The examples are taken from the C-ORAL-BRASIL I corpus (Raso & Mello, 2012).

The first context is the nucleus – or part of the nucleus – of the illocutionary unit (that is, the nucleus of COM). In this context, ASSIM can be paraphrased with its full semantic meaning (like this):

ASSIM in illocutionary nucleus of COM

Audio file 20 - bfamcv04_191-196

*BRU: [191] *cê pode fazer **assim** //*

*BRU: [192] *que isso é <similar> //*

*HEL: [193] *<tá> //*

*HEL: [194] *e **assim** //*

*BRU: [195] *não //*

*BRU: [196] ***assim** //*

BRU: [191] you can do it **this way //*

**BRU: [192] 'cause this is <similar> //*

**HEL: [193] <ok> //*

HEL: [194] and **this way //*

*BRU: [195] *no* //
*BRU: [196] *this way* //

The three instances of ASSIM are replaceable by “in this way / like this” and are sufficient for the illocution to be conveyed. ASSIM can also occur in a dedicated non-illocutionary IU. The informational functions assumed in this case can be both textual, or interactional. First, I show one textual function. The following example illustrates an utterance in which ASSIM assumes the semantic function of modalization:

ASSIM in a dedicated unit with a textual function

Audio file 21 - bpubcv03_123

FER: [123] pra gente nũ ter uma tradução <bem> / &he / chula / **assim** / bem
ao pé da letra horrorosa / aí fica <complicado> //

FER: [123] *so that we don't end up with a translation very / &he / pimp / let's
say so / very literal and poor / this gets <complicated> //*

In this example, ASSIM can be paraphrased by “let's say so” with the intention of attenuating “pimp”. It thus assumes a function compatible with that of a modalizing Parenthetical (PAR), i.e., a textual IU.

In addition to textual functions, ASSIM can take on interactional functions typical of DMs in the sense of this research. The following example shows a case where ASSIM no longer turns to the text of the utterance but to the interaction itself:

ASSIM in dedicated unit with interactional function

Audio file 22 - bpubdl05_254

GET: [254] então / é uma abelha que / **assim** / também tem um futuro como
polinizador / né //

GET: [254] *so / this is a bee that / you know / also has a future as a pollinator
/ huh //*

I will talk about the specific function of this realization later on. But it can be said that the attitude with which the assertion is enacted does not leave much place for interpretations such as “like this,” or “let's say so” in the same approximative way it is used in the Parenthetical. Here, the speaker seems to use this unit to draw the addressee’s attention to a conclusion, i.e., to the point they were trying to reach to. I do not dive into the prosodic differences of the examples given, but their coherence with information functions is pointed out in Raso & Santos (2020).

By showing these examples, I tried to show how the same lexical item can assume different informational values depending on the prosodic realization; one of the basic assumptions of the L-AcT is that each IU is correlated with a prosodic form that guides, at the forefront, the interpretation of its informational value. The point of these examples is to show that the prosodic segmentation is not enough, notwithstanding its importance. Together with distributional constraints and the prosodic segmentation, the prosodic realization of a specific content will serve as a formal criterion enabling us to analyze and discriminate between illocutions and DMs. The examples given in this section will now help us respond to the poly-functionality of DMs.

3.2.3 Poly-functionality of DMs

Our view agrees with the statement that the same lexical item or small expression can take on different functions. However, the same does not hold true for the statement saying that a concrete item (an occurrence/token) can bear more than one function at the same time. At least not if we mean function of the same level.

As explained in Chapter 2, the L-AcT put forth the hypothesis according to which there is an isomorphic relationship between the prosodic unit and the Information Structure (with the exceptions explained therein). Furthermore, different information functions seem to be correlated with specific prosodic realizations. The relationship

established by the different informational functions is of paradigmatic nature. When a unit is TOP, it cannot be COM at the same time. As a matter of fact, one can observe some prosodic variation among different realizations of the same function. But this variation seems related to different attitudes – as defined in Mello & Raso, (2011) and Raso & Rocha (2016) –, to emotions, as well as to many other sociolinguistic variables. A prosodic form will convey a unique interactional function. For analytical purposes, when there is doubt between two functions, we can also resort to distributional constraints.

3.2.4 Functions of DMs

Thus far, I have explained why prosody must be given a primary role on the study of DMs. Now I would like to deal with some of the functions described in the literature. I begin with the most obvious. DMs should not be confused with illocutions; not in the sense that DMs are dealt with in this research. Illocutions are textual units that build the semantic content of the utterance; illocutions carry the speech act being enacted by the terminated sequence. DMs are not semantically compositional with the text of the utterance.

A second function mentioned in the literature is related to the notion of *modality*. Assuming the sense given by Bally (1950), i.e., that the modality marks the position taken by the speaker with respect to the expressed content, we have to admit that there is a semantically compositional relationship between modal operators and the propositional content of the utterance. By attributing modal functions to DMs, we call into question the premise that there is no compositionality between the DM and the utterance. If the unit is compositional (and thus build the text of the pattern), it is not a DM in the sense dealt with in this research. This could be sufficient to say that DMs do not take on modal functions, especially when we consider that DMs are also desemanticized.

3.2.5 Summary of the section

The L-Act offers analytical criteria whereby one can isolate DMs from other kinds of units; these principles are of prosodic and informational nature. The prosodic segmentation gives a first cue about the non-compositionality of a unit. The prosodic form works as the formal principle that, at the forefront, enables the distinction of different types of informational functions. As a matter of fact, lexical and syntactic formal features do not allow for the identification of DMs. The lexicon can take on any function depending on the concrete realization. The lexicon is also variable over time and space, whereas prosody is more stable. We do not deny that many factors and aspects play a role in assigning subfunctions to DMs within context. What is argued is that prosody plays a leading/mapping role in this assignment. The prosodic form works as a primary branching mechanism that, together with other aspects, leads to a specific, contextualized subfunction. The prosodic form is stable, while the lexicon can vary greatly. The subfunction can be determined by the lexicon, through the interaction of the lexical item with prosody and with the context in which the DM is produced. In the subsections to follow, I present our last proposal for the macro-functions of interactional DMs. As much as possible, I try to match the functions indicated in the literature with the functions offered by our proposal, showing how that macro-function and prosodic form are coherent with proposed subfunctions.

3.3 L-ACT'S DISCOURSE MARKERS FRAMEWORK HISTORY

This concise presentation offers a historical overview of the examination of DMs within the Language into Act Theory (L-Act) research framework. The latest proposal identifies and addresses various issues in the preceding descriptions.

The initial proposal, put forth by Cresti (2000), introduced four DMs (Dialogic Units in L-Act's terminology): Incipit (INP), Conative (CNT), Allocutive (ALL), and Phatic (PHA). According to Cresti's proposition, the INP's function involves initiating the turn or utterance while expressing affective contrast with the preceding utterance. It

consistently occurs at the outset of an utterance or at the beginning of the sub-pattern of a stanza. Prosodically, INP displays high f₀ with respect to illocution. Cresti (2000) observed three f₀ profiles: rising, falling, and rising-falling, without addressing the reason for the apparent variability in conveying the same function.

CNT, described by Cresti (2000), aims to persuade the addressee to undertake or cease a specific action. Its distribution is unrestricted. Prosodically, CNT features a falling f₀ profile, short duration, and elevated intensity, though not as high as those observed in INP.

ALL serves the purpose of establishing social cohesion among conversation participants or clarifying the utterance's addressee. Typically filled with titles, epithets, and proper names, ALL exhibits a free distribution according to Cresti's proposal.

As per Cresti's initial proposal, PHA aims to keep the communication channel open. Despite being noted for its short duration and low intensity, Cresti did not assign a specific f₀ profile to this unit. Additionally, PHA is suggested to have a free distribution, contributing to challenges addressed later in this work.

Finally, EXP, according to Cresti's proposal, expresses emotional support for the illocution and exhibits a distributionally free nature. Prosodically, EXP displays mean intensity and duration, with an f₀ profile described by the author as modulated, allowing for one or more f₀ movements.

Raso (2014) attempted an initial systematization of Cresti's (2000) framework by comparing DM samples from the Italian and Brazilian Portuguese C-ORAL corpora. Recognizing the need for a refined prosodic description, Raso (2014) laid the groundwork for subsequent works. Raso & Vieira (2016) addressed apparent variations in INP's f₀ contours and partly elucidated the prosodic distinction between CNT and ALL. Gobbo (2019) introduced a supervised classification model focusing on parameters derived from prosodic-acoustic measurements for the three most clearly defined functions: INP, CNT, and ALL. His model achieves a 0.86 goodness of fit (as measured by the accuracy score) with eight prosodic parameters. Raso

et al. (2022) further delved into L-Act's DM framework, establishing methodological foundations for this research. The subsequent section presents our most recent proposal. At this point, I would like to reflect on the problems to be tackled:

- a) What is the function of the PHA? Keeping the channel open can be attributed to many things, starting with TMT. What would be the prosodic correlates of PHA?
- b) What does emotional support mean in EXP? It does not have a defined form either.
- c) How can the formal variability of INP be explained?

With unexplained variable forms (EXP, PHA and INP) it would not be possible to guarantee that the prosodic form is the formal vehicle of the function.

3.4 THE MOST RECENT PROPOSAL FOR DMs

This proposal encompasses five Discourse Marker (DM) units. The two most clearly defined units from prior research (CNT, and ALL) have been kept unchanged from Raso & Vieira (2016) and Raso & Ferrari (2020), as their descriptions were considered satisfactory due to the coherent mapping between prosodic form and assigned functions. Functionally, we have redefined EXP and assigned a specific prosodic form to it. The Phatic Unit was excluded. According to Cresti (2000), PHA's function is to maintain an open communication channel. However, this function can be accomplished by various other devices, such as filled pauses or Scanned Units. PHA lacked a distinct function, its prosodic form was somewhat unspecified, and its distribution did not contribute significantly to disambiguation. Furthermore, in first moment we hypothesized a new form tentatively called FLAT. Then, we

observed that the FLAT and INP shared a distinctive common trait: both exhibit a clear flat contour over the stressed syllable. But INP had a high f₀ profile and FLAT a low f₀ profile. This variation can be better accounted for as function of the absence or presence of an attitude of contrast with respect to what was said before. Thus, the description of INP presented in Raso & Vieira (2016) was also received but with modifications. One new unit has been introduced and is presented subsequently.

In the revision process leading to this proposal, three key steps were undertaken:

- a) a reassessment of prosodic boundaries annotation;
- b) an exploration of regularities in prosodic forms;
- c) an examination of these regularities in the context of previously identified and newly proposed interactional functions.

The ensuing section outlines the five DM functions. A comprehensive summary table, featuring functions, forms, distribution, and frequency of each DM, is provided at the conclusion of this proposal.

3.4.1 The Incipit (INP)

In the preceding subsection, attention was directed towards the diverse f₀ profiles described by Cresti (2000) for the INP unit. As elucidated by Raso & Vieira (2016), INP can exhibit a remarkably wide f₀ range within an exceptionally brief duration. Its prosodic form is characterized by a flat f₀ profile over the stressed vowel, accompanied by very high intensity level—exceeding the mean value of COM in both instances.

There are two types of INP. The first one is the high INP, which

marks a contrast with respect to what was said before. This type is characterized by a high flat f0 profile over the stressed vowel. When the stressed vowel is preceded by voiced segmental material, a rising f0 movement is observed before reaching the higher value of the stressed vowel. Similarly, when the stressed vowel is followed by voiced material, the profile assumes a falling f0 movement after the stressed vowel. In instances where both rising and falling movements are present, the highest f0 level is at the stressed vowel. It is worth noting the perceptible tenseness on the stressed vowel of INP, although this aspect remains outside the scope of evaluation in the current work. The following examples illustrates these characteristics:

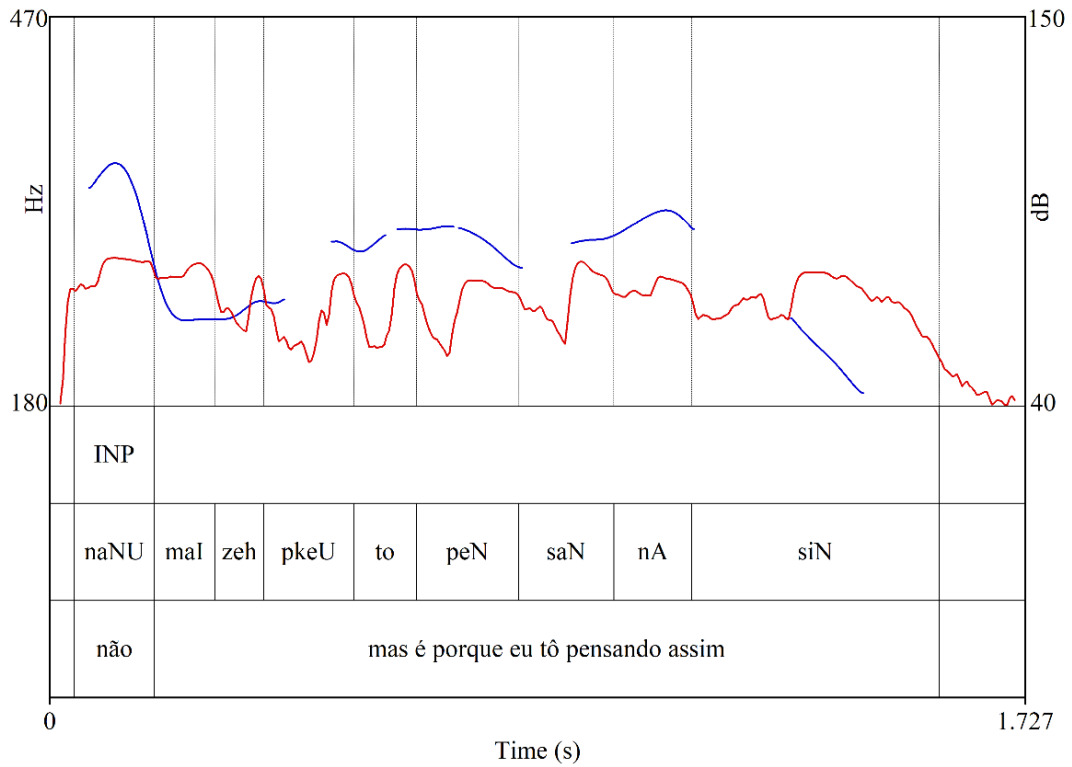
Example of a high INP

Audio bfamdl02_197

*BAL: não /=INP= mas é porque eu tô pensando assim //

no / but it is because I'm thinking this way //

Figure 9 - Form of high INP



The figure 9 above shows how the form of the INP is affected by an initial voiced consonant that causes a rising movement until the vowel of the diphthong, which sharply falls in the semivowel.

Another type of INP is the one that does not mark a contrast to previous content (flat INP). It has intensity and duration similar to the contrastive INP. But the flat profile is low with respect to COM, as in the example and Figure 10 below:

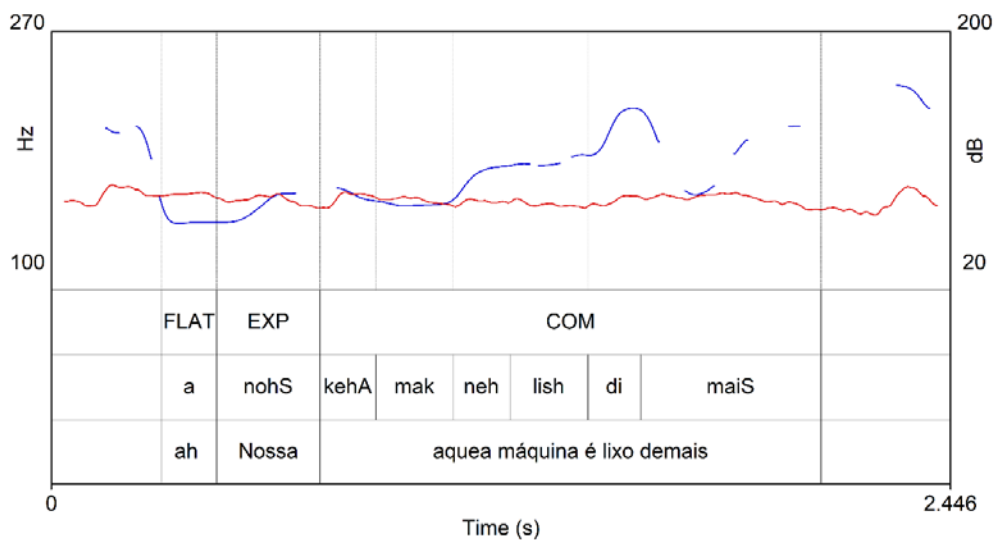
Example of a flat INP

Audio file 23 - bfamnn05_102

*JUN: ah /=FLAT= Nossa /=EXP= aquea máquina é lixo demais // =COM=

oh /=FLAT= Holy / that camera is complete trash //

Figure 10 - Form of flat INP



3.4.2 The Conative (CNT)

CNT is distinguished by a falling f_0 movement, often accompanied by a high f_0 variation rate. However, the variation rate may be influenced by attitudinal factors. This fall is not as pronounced as the movement seen in INP outside the stress and is lower than the mean f_0 of COM. In contrast to ALL, where the f_0 movement falls from the unit's onset, the CNT's falling movement aligns with the stressed vowel. Raso & Ferrari (2020) noted that a slightly rising f_0 movement (a preparation) can be observed in the presence of voiced segmental material before the stressed vowel. This preparation is more noticeable when the stressed syllable is not initial, but it may also be discerned when there is sufficient voiced material before the stressed vowel. Raso & Ferrari (2020) proposed a more precise functional definition of CNT,

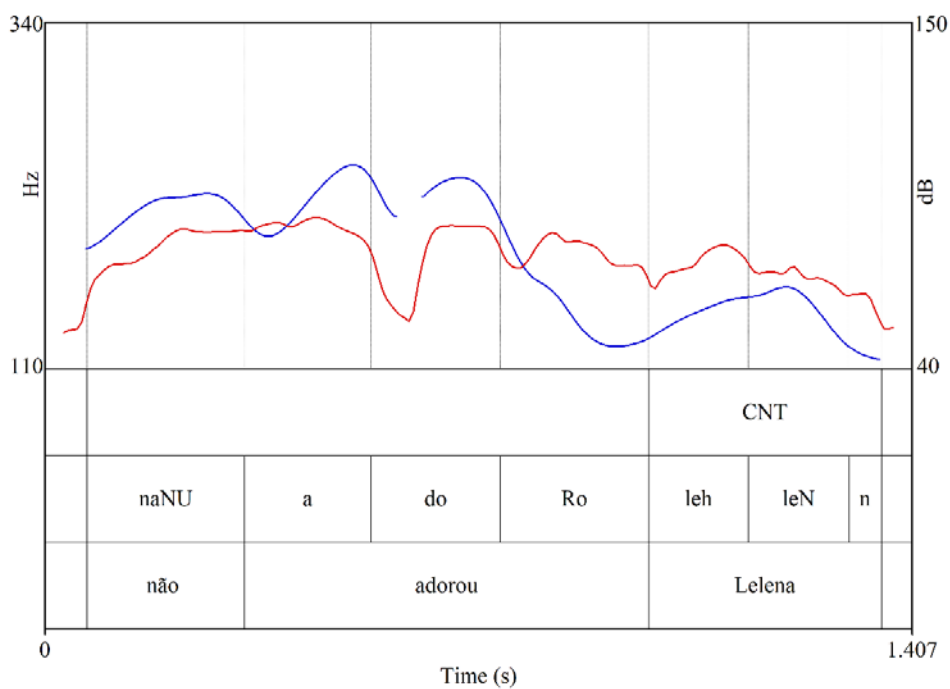
indicating its role in signaling the illocutionary resolution of the utterance.

Example of CNT

Audio file 24 – btelptv06_094

*LUR: não / adorou / Lelena // = CNT =
no / he liked it / Lelena //

Figure 11 - Form of CNT



We can observe in Figure 11 the alignment of the falling f0 movement with the stressed syllable and the rising movement over the pre-stressed syllable.

3.4.3 The Allocutive (ALL)

Cresti (2000) proposed the occurrence of ALL in any position. However, upon closer examination of this unit's behavior, Raso & Ferrari (2020) observed that it does not manifest at the beginning of a pattern and tends to favor the final position. Although very few instances were identified in medial positions, these occurrences provided valuable insights into the f₀ profile of ALL. Irrespective of the stress structure of the lexical item, the f₀ profile of ALL falls along the unit's onset and then flattens. ALL exhibits lower intensity levels than other units and some segmental lengthening. Whereas the portions under the falling f₀ movement are phonetically well articulated, segments of the flat portion often undergo phonetic reduction, and their intensity may be so low that f₀ estimation becomes challenging.

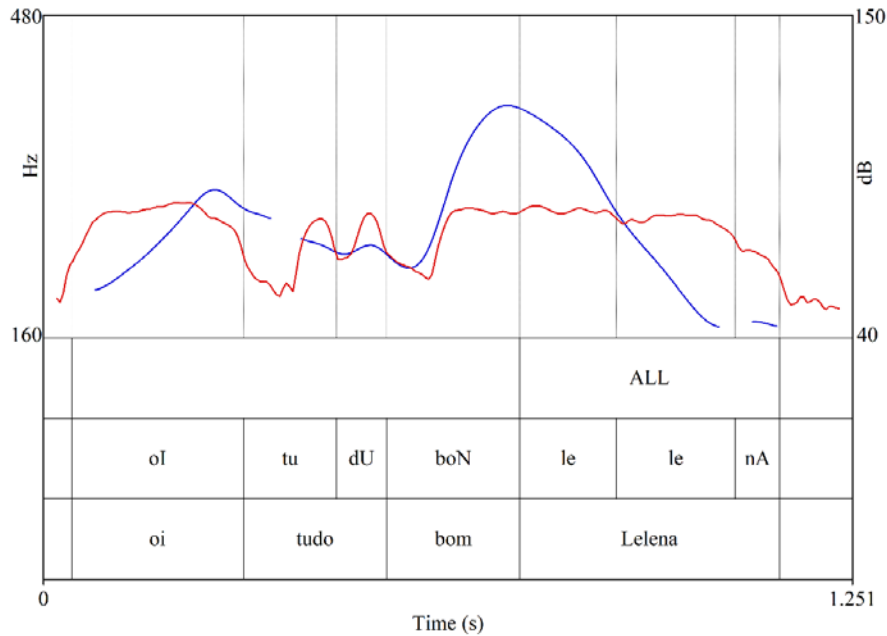
Example of ALL

Audio file 25 – btelpv06_003

*LUR: oi / tudo bom / Lelena // =ALL=

hello / everything OK / Lelena //

Figure 12 - Form of ALL



3.4.4 The Expressive (EXP)

EXP is, according to previous works, utilized to express emotional support for the illocution, highlighting its distributional freedom and the potential inclusion of multiple f₀ movements in its prosodic form. However, certain issues arise from this characterization. The definition of providing emotional support to the illocution is overly vague, and furthermore, a distinct prosodic form has not been assigned to the unit.

We maintain that there is a specific DM that we still call EXP, but that we describe in a clearly different way. EXP is employed to convey surprise but is enacted in a manner that prevents it from being interpreted as an illocution. It manifests a rising f₀ movement until the stressed vowel, which may briefly fall in the presence of segmental material after. There is a marked lengthening on the stressed syllable with respect to the mean syllabic duration of COM. The intensity of EXP is comparable to that of COM or slightly lower. This unit

consistently appears at the beginning of the pattern.

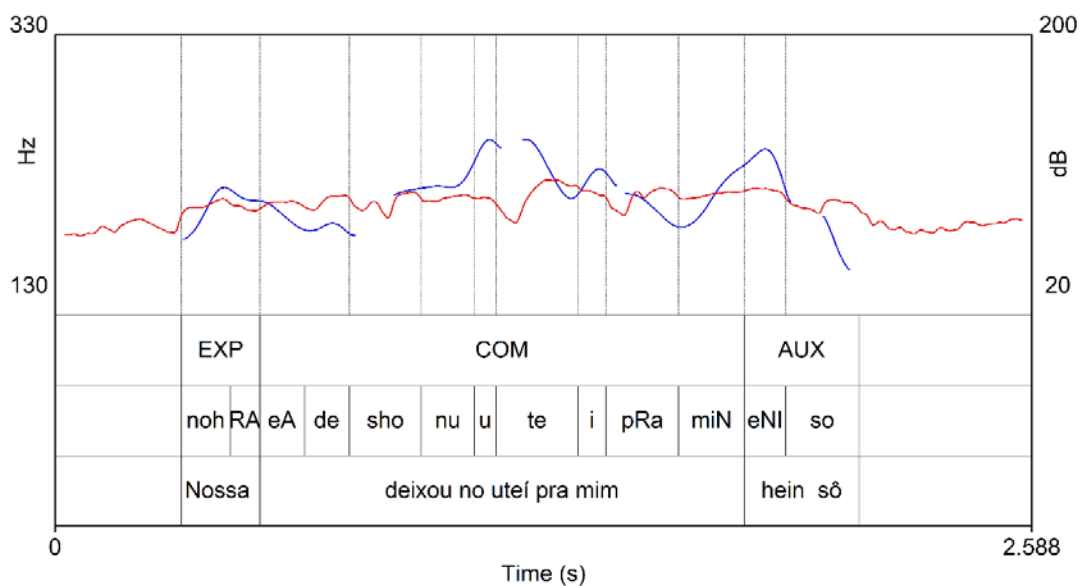
Example of EXP

Audio file 26 – bfammn05_102

*TON: Nossa /=EXP= ea deixou no uteí pra mim /=COM= hein sô //CNT=

Holy /=EXP= she left it in the UCI for me / you saw //

Figure 13 - Form of EXP



3.4.5 The Highlighter (HGL/EVD)

The Highlighter was provisionally labeled as EVD (Evidenciador in BP). It directs the addressee's attention to the preceding statement, often performed with a focus. HGL/EVD is typically produced with a slightly rising f0 movement and significantly lower intensity than the COM. The duration is considerably shorter than that of COM. The slope of the rising movement can range from nearly flat to distinctly rising,

contingent on the speaker's attitude.

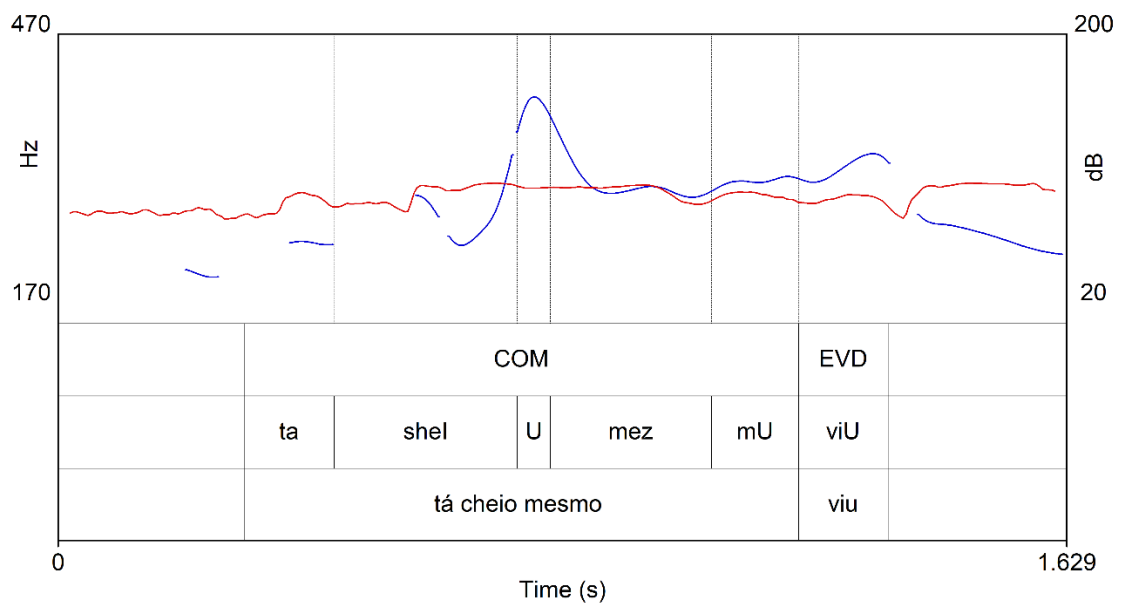
Example of EVD

Audio file 27 - bfamdl01_201

*REN: tá cheio mesmo /=COM= viu //EVD=

it's really crowded /=COM= huh //EVD=

Figure 14 - Form of EVD

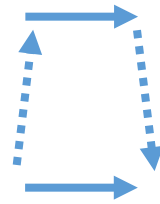
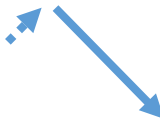





3.4.6 Summary table

The table provided below offers a summary of the functions, fundamental f0 movements, intensity levels, durations, and distributions of Discourse Markers (DMs) in relation to the illocutionary unit. Dashed lines represent non-mandatory f0 movements. Additionally, the intensity and duration levels are indicative trends that require a more refined statistical description. The dataset's token count

for each identified function is also included.

Table 4 - DM summary table

UNIT	FUNCTION	BASIC F0 MOVEMENT	INTENSITY	DURATION	DISTRIBUTION
INP	Begin the utterance		Higher than COM	Much shorter than COM	Beginning of the pattern
CNT	Point to an illocutionary solution		Lower than COM	Shorter than COM	Free
ALL	Establish social cohesion		Lower than COM	Shorter than COM	Middle or at the end of the pattern
EXP	Convey surprise in a non-illocutionary way		Paired to COM	Paired to or longer than COM	Beginning of the pattern
EVD	Highlight what was said		Lower than COM	Shorter than COM	Middle or at the end of the pattern

The ensuing section will delve into the materials and methods employed in this research, encompassing a description of the corpora utilized, the methodologies applied to model prosodic form.

4 MATERIALS AND METHODS

This chapter details the materials and methods employed to substantiate the recent proposal presented in Chapter 3. This chapter is centered on the prosodic-acoustic features utilized for modeling the prosodic forms of Discourse Markers (DMs), specifically delving into three distinct feature groups: intensity, speech rate, and fundamental frequency (f_0). Firstly, I delineate the corpus used for this work, specifying its characteristics. Then, I give the details of how the DM sample was obtained and annotated. I also present the procedures utilized for the standardization of prosodic features. Finally, I present the prosodic features used for the DM description and the classification task. All scripts and notebooks used in this research are available at <https://github.com/saulo-smendes/discourse-markers-scripts>. Other materials, both for the statistical and experimental analyses, are available at <[SHARED MATERIALS THESIS](#)> or via the QR code below:



4.1 C-ORAL CORPUS

4.1.1 Core characteristics

The analysis of the prosodic forms of DM was based on a sample extracted from the spontaneous speech corpus C-ORAL-BRASIL I (Raso & Mello, 2012), a corpus representing spontaneous spoken Brazilian Portuguese, especially from the diatopy of Minas Gerais. The C-ORAL-BRASIL was structured to be comparable to the corpora of the C-ORAL-ROM family (Cresti & Moneglia, 2005), representing French, Italian, Spanish and European Portuguese. For this work, the most important feature of the corpus is its annotation with prosodic information (Cresti & Moneglia, 1997). The corpus is annotated with terminal boundaries, delimiting utterances and stanzas (terminated sequences), and non-terminal boundaries, delimiting prosodic units.

Another important feature of the corpus is that it recorded spontaneous speech in natural and wide diaphasic contexts, i.e., situational variation, unlike controlled situations, in which linguistic behaviors are highly predictable (Raso & Mello, 2014). Situational variation generally entails actional variation, i.e., variation in the linguistic actions being performed (illocutions). Variation in linguistic actions is, in turn, a decisive factor in recording a greater number of speech structures, as variation at the level of the Information Structure (Raso & Mello, 2014). With more varied structures, one can observe more contexts where the same lexeme can occur with different informational values, which is crucial for studying DMs.

In general, the corpora of the C-ORAL family are provided with special features enabling multilevel research. All corpora are comprised of: a) audio files, textual files, text-to-speech alignment files supported by the software WinPitch (Martin, 2015); b) metadata; c) lexical and morphosyntactically tagged textual files¹²; d) frequency

¹² Particularly, the C-ORAL-BRASIL corpora were tagged with the parser PALAVRAS (Bick, 2012).

lists, measurements of the corpora, and statistical data of participants.

The textual format of the C-ORAL corpora followed the CHAT Transcription Format (MacWhinney, 2000) adapted for segmentation (Moneglia and Cresti, 1997). The format encompasses two levels. On the first level, headers contain the metadata of the recorded event, such as participants' socioeconomic background, topic, situation, and place of the recording session, as well as information on the audio file, such as word count, length, and acoustic quality classification. The second level contains the transcriptions of speech, paralinguistic and nonlinguistic events, and the segmentation. The text can be organized by turn or terminated sequence, which we introduce in the next subsection.

To enable the study of the informational structuring of speech, the corpora of some of the project's languages have also been equipped with informationally annotated minicorpora, following the architecture of the matrix corpora (see Martínez et al., 2018; Panunzi & Gregori, 2012; Panunzi & Mittmann, 2014). In addition, the project also has other linguistic resources, already compiled or being compiled, such as for example, the American English minicorpus (Cavalcante & Ramos, 2016, with texts extracted from the Santa Barbara Corpus of Spoken American English - SBSCAE, Du Bois et al., 2000), C-ORAL-ANGOLA (see B. Rocha et al., 2019, for details of the progress of the compilation) and C-ORAL-ESQ (Ferrari et al., in preparation).

4.1.2 Organization of the C-ORAL-BRASIL corpus

The C-ORAL-BRASIL corpus was organized into four sub-corpora. These sub-corpora are:

- I. Subcorpus of Informal Speech in a Natural Context
- II. Subcorpus of Formal Speech in Natural Context
- III. Media Subcorpus

IV. Subcorpus of telephone calls

In addition, the work of compiling and publishing the corpus was divided into two stages. The first stage was completed with the publication of the Informal Speech in Natural Context subcorpus, C-ORAL-BRASIL I, in 2012 (Raso & Mello, 2012). Covering the Formal Speech in Natural Context, Media, and Telephone Calls subcorpora, the second stage, C-ORAL-BRASIL II, has already completed its compilation and validation work, and its results will soon be published (Raso, Mello & Ferrari, in preparation). Only the C-ORAL-BRASIL I was sampled for this study since it has already been published and is available for research.

The texts in C-ORAL-BRASIL I were divided into two social contexts: family/private and public (Raso, 2012a). The texts were organized by interactional typology. The monologues included texts in which speech is predominantly monologic, i.e., carried out mostly by just one participant. In dialogues, the interactions are more evenly distributed between two participants. Finally, in conversations, three or more participants interact. The architecture of the C-ORAL-BRASIL subcorpora are presented in Table 5 and Table 6 below:

Table 5 - Informal subcorpus (C-ORAL-BRASIL I)

Language register	Social context	Structure of the communication event	Number of words	Number of files
Informal	Family/private	Monologues / dialogues / conversations	159,364	105
	Public	Monologues / dialogues / conversations	48,766	34
1. Informal in Natural Context (Subtotal)			208,130	139

Table 6 - Subcorpora and domains of use of the C-ORAL-BRASIL II

Language register	Subcorpus	Domain of use	Number of words	Number of files
Formal	Natural Context	Business	10,851	4
		Conference	17,320	9
		Law	16,107	9
		Political debate	15,707	12
		Political speech	16,047	15
		Preaching	12,826	9
		Profession explanation	16,247	8
		Teaching	16,291	8
Subtotals (Natural Context)			139,647	74
Formal	Media	Documentary	23,530	29
		Extra	24,728	16
		Interview	15,506	9
		Meteorology	232	1
		News	6,096	9
		Scientific Press	13,233	12
		Sport	12,234	7
		Talk show	44,088	18
Subtotals (Media)			121,396	101
Informal	Telephone	Private conversations	25,533	50
		Public conversations	5,755	29
Subtotals (Telephonic corpus)			31,308	79
Totals (CORAL-BRASIL II)			292,351	254

Together, the C-ORAL-BRASIL I and the C-ORAL-BRASIL II make up a total of 393 files and 500,481 words, thus being a medium-sized spontaneous speech corpus. With respect to the time period represented, the vast majority of recording sessions were carried out between 2009 and 2017.

4.1.3 Segmentation

4.1.3.1 Annotation scheme

Texts were prosodically parsed into terminated sequences, signaled by a terminal boundary transcribed with a double-slash sign (/), and non-terminal prosodic units marked by a non-terminal boundary transcribed with a single-slash sign (/).

Example of segmentation

Audio file 28 - bnatte03_093-094

*ALA: [93] como é que ele vai saber lá / qual foi a parte do texto <que tinha> +

*GER: [94] <eu te falei> / ele tem / internamente aqui / **alguns tipos de [/3]** como posso dizer pra vocês / certos parâmetros / que ele vai tirar / do texto //

*ALA: [93] *how was it supposed to recognize / which part of the text <was to be> +*

*GER: [94] *<I told you> / it has / internally here / **some kinds of [/3]** how can I put that to you / certain parameters / that will be drawn / from the text //*

In the example above, the speech stream between the beginning of the speaker's turn and the terminal boundary sign forms a *terminated sequence*. However, the speaker's whole turn may be formed by multiple terminated sequences, as in the example below, where digits in square brackets signal the beginning of another terminated sequence. Texts are aligned to speech on the terminated sequence level.

Example of segmentation

Audio file 29 - bnatbu02_001-002

*NEU: **[1]** aqui o' // **[2]** eu tenho ele com braço de vinte-e-cinco / e tenho e' com braço de quinze //

*NEU: **[1]** take a look // **[2]** I have the twenty-five [width sofa] arm option / and I have the fifteen option //

The segmentation also signals retractions accompanied by non-terminal boundaries. Broadly, retractions are the withdrawal of a part of the text often triggered by changes in the speaker's initial speech program or by mistakes. They are also frequently used as a time-taking device in which case a word may be repeatedly uttered. In case retractions trigger a non-terminal boundary, the boundary sign is followed by a digit that indicates the number of words retracted. This sign is embedded in square brackets. In the first example, the retraction of *alguns tipos de* (some kinds of) is signaled by *[/3]*.

Speech may also be interrupted before getting to a point where a terminal or non-terminal boundary would be properly or completely signaled. This occurs when speakers abandon their program either by starting another turn or as a consequence of having their turn taken by another participant. Interruptions are indicated by a plus sign (+), as shown in the first example.

4.1.3.2 Validation of the segmentation

The segmentation underwent a specific validation, which is out of the scope of this work. The methods and results are further detailed in Mello et al. (2012). It is nonetheless important to present them here. The prosodic boundary annotation was validated by measuring the

reliability of the inter-annotator agreement on the annotation of prosodic boundaries. Two validations were carried out, one before the bulk of texts were prosodically parsed and another after compilation, just before the corpus final revision. These tasks aimed not only at standardizing the annotation beforehand but also at ensuring its quality at the last compilation step.

The degree of agreement was evaluated by the Fleiss' kappa statistic (Fleiss, 1971), which assesses the reliability of agreement between more than two raters in assigning categorical classes. The degrees of agreement that should be met were established during the planning phase of the corpus. It was established that the kappa values of at least 0.8 (almost perfect agreement) for terminal boundaries and 0.6 (substantial agreement) for non-terminal boundaries should be met.

The assessment of the degree of agreement was done as follows. Each annotator received texts (audio plus transcripts without prosodic boundary signs) to be annotated within the following three days. The task encompassed texts of mostly monologic and dialogic interactions. Each word boundary was a candidate for receiving the boundary sign. If there was no boundary, the blank space should not be changed. If annotators perceive a boundary, they should indicate it by adding the proper sign for a non-terminal or terminal boundary. Because they are special cases, retraction and interruption signs were left aside. For instance, retractions may trigger (but not always) non-terminal boundaries.

Three different agreement rates were calculated. The overall agreement was obtained by adding all possible positions (i.e., all word boundaries) and considering the agreement on the absence of a boundary, the presence of a non-terminal boundary, or the presence of a terminal boundary. In its turn, the partial agreement was aimed at putting in evidence how salient a boundary of any kind was. To calculate it, all positions were considered, as in the overall agreement, but this time, no distinction between the two boundary types was made. Finally, in a more conservative approach, the realistic agreement was calculated by tallying only the positions where at least one

boundary of any kind was marked, thus eliminating the effect that word boundaries without perceived prosodic boundaries would have over the statistic.

After the agreement rates were calculated, the divergences were discussed in each group. After each session, another annotation task was repeated until the kappa values of 0.8 (terminal) and 0.6 (non-terminal) for the overall agreement were reached. It is noteworthy that both groups always displayed good agreement rates for terminal boundaries and that, as pointed out by Moneglia et al. (2010) and Mello et al. (2012), experience and practice played a crucial role in the recognition of boundaries. The results of the partial agreement showed that annotators had no problem distinguishing between the absence and presence of boundaries (kappa values always higher than 0.84 for the best-performing group and 0.75 for the least-performing group). Divergence was recorded mostly for the distinction between non-terminal and terminal boundaries.

The kappa values for the realistic agreement were met in the validation done after the first compilation phase, just before the corpus underwent its final revisions. The kappa values achieved by the group in charge of revising the segmentation are shown in the table below. They show that, even with the most conservative approach, prosodic annotators achieved

Table 7 - Kappa values for the realistic agreement rate before segmentation validation

Type of agreement	Total	Dialogues	Monologues
Realistic agreement	0.65 (substantial agreement)	0.66 (substantial agreement)	0.63 (substantial agreement)
Terminal boundaries	0.81 (almost perfect agreement)	0.80 (almost perfect agreement)	0.80 (almost perfect agreement)
Non-terminal boundaries	0.62 (substantial agreement)	0.65 (substantial)	0.59 (moderate agreement)

Type of agreement	Total	Dialogues	Monologues
-------------------	-------	-----------	------------

agreement)

(Adapted from Mello et al., 2012: 165)

This procedure was maintained in the compilation of the C-ORAL-BRASIL II by having the most experimented and best-performing prosodic annotators on the upper-level tasks. To be responsible for the segmentation revision, a group of annotators must have achieved an overall agreement rate of 0.80 (for terminal boundaries) and 0.60 (for non-terminal boundaries), as measured by Fleiss' kappa values.

Considering the important consequences of prosodic parsing over syntax and information structure, the Laboratory for Empirical and Experimental Linguistic Studies (LEEL/UFMG) in partnership with the Phonetics Laboratory of the Campinas University has been carrying out research aimed at identifying the acoustico-phonetic features guiding the production and perception of boundaries and at developing models for their automatic detection on spontaneous speech (for further information see: Barbosa and Raso, 2018; Raso, Teixeira, and Barbosa, 2020; Teixeira, Barbosa, and Raso, 2018). The studies have examined a large number of acoustic measurements extracted from a time window positioned around prosodic boundaries of the same type marked by at least 50% of annotators. These measurements encompass speech rate and rhythm, standardized segment duration, fundamental frequency (f0), intensity, and silent pauses. Although it is still ongoing work, the results obtained to date are promising. To this point, the accuracy of the best models is at 0.74 for terminal boundaries and 0.66 for non-terminal boundaries in a cross-validation set. These models show that silent pause and f0 features are the most important elements contributing to terminal boundaries detection and, on the other hand, that duration plus pause features contribute to the best-performing non-terminal boundary model. Studying the role played by pauses is presently an issue of major concern since they have also been shown to be a confounding element in the distinction between non-terminal and terminal boundaries.

4.1.4 Text transcription

The transcription of the textual content followed an orthographic-based norm coupled with a set of special criteria (Raso & Mello, 2009, 2012). The orthographic norm enables texts to be easily understood and handled by users. It also enables text to be automatically processed without dealing with an uncountable number of variant forms unknown beforehand, such as partial or full phonetic-based transcriptions. However, many characteristics of spontaneous speech cannot be overlooked. The set of special criteria aimed, thus, at documenting possible grammaticalization and lexicalization processes ongoing in the language. Without this documentation, many phenomena deserving further investigation would be lost, or their study would be more complex. The set also established the transcription signs for several other non-linguistic and paralinguistic phenomena of pragmatic interest. The transcription criteria are available in Appendix A (11).

4.1.5 Morphosyntactic parsing

The C-ORAL-BRASIL corpus also contains morphosyntactic annotation files. These files present the syntactic functions of each word, as well as their respective morphological classification. These elements allow advanced searches on observable syntactic patterns in speech. The annotation was done using the PALAVRAS morphosyntactic annotator (Bick et al., 2012).

4.1.6 Minicorpus

To conduct research on Information Structure according to the L-Act, identifying prosodic units is a preliminary step. However, annotating these units is a laborious task that requires a considerable amount of time and collaboration from trained individuals. Corpora like those of the C-ORAL family can hardly receive a full informational annotation. For instance, the informal part of the C-ORAL-BRASIL corpus alone

contains nearly 62,000 prosodic units (Cavalcante, 2020). To address this issue, scaled-down versions of these corpora have been created, which correspond to the minicorpora referred to at the beginning of this chapter. The minicorpora of the C-ORAL family maintain the same architecture than the matrix corpora from which they derive. The BP minicorpus comes from the C-ORAL-BRASIL I corpus, and it is composed of an equivalent proportion of monologues, dialogues, and conversations. The minicorpora are also somewhat balanced in terms of the distribution of their texts according to sociological context (family/private and public). Besides the BP minicorpus, informationally annotated minicorpora are also available for Italian (Cresti et al., 2022), American English (Cavalcante & Ramos, 2016), and Spanish (Martínez & Somacarrera, 2018).

4.1.7 Availability

The C-ORAL-BRASIL I (both the matrix corpus and the minicorpus) is fully available for download at <www.c-oral-brasil.org>. In addition, the minicorpora can be consulted through the Database for Corpora Multimedia platform (DB-CoM, available at <<http://www.c-oral-brasil.org/db-com>>) and the Database for Information Patterning Interlinguistic Comparison, the DB-IPIC (<<http://www.lablita.it/app/dbipic/>>).

4.2 DM SAMPLING

A total of 1025 tokens were annotated as a type of Dialogic Unit (Discourse Marker) in the BP Minicorpus. This sample was revised in Gobbo (2019), resulting in the following distribution across the pragmatic functions that were attributed to them:

Table 8 - DM tokens in the BP minicorpus (Gobbo, 2019)

Discourse	DCT	ALL	CNT	INP	AUX	Total
-----------	-----	-----	-----	-----	-----	-------

Marker ¹³						
BP minicorpus	173	63	84	40	665	1025
Not analyzed	173	12	15	4	261	465
Analyzed	0	51	69	36	404	560

According to Gobbo (2019), many observations had to be discarded because it was impossible to apply the acoustic extraction procedures or because the tokens were unreliable. The main motives leading to discard a unit were: (i) overlapping speech or too much background noise; (ii) token could not properly be segmented; (iii) f0 curve could not be adjusted acceptably; (iv) utterance as a whole was unreliable; or (vi) absence of reference unit (see 4.4.1. Reference for the standardization of prosodic-acoustic parameters).

During his research, Gobbo's model (2019) accounted for the three DM functions accurately labeled in the BP minicorpus. However, the author points out that some data were deliberately not discarded rigorously to preserve a sufficient quantity of observations for data analysis. Gobbo's (2019) sample was further revised here. This revision aimed to check audio quality issues and the existence of prosodic boundaries (see Santos & Raso, 2022, for some biases in speech segmentation) and establish a proper categorization of tokens under the AUX label. The results of this revision are shown below Table 9.

Table 9 - Discarded tokens by criterion

Discard criterion	Number of tokens
Absence of boundary	160
Overlapping speech	1
Other quality issues	17
Other reasons	10
Illocutionary	64

¹³ DCT: Discourse Connector; ALL: Allocutive; CNT: Conative; INP: Incipit; AUX: tag is used to label information units that probably are Discourse Markers but whose functions have not been identified.

Discard criterion	Number of tokens
Total	252

From Gobbo's (2019) work, a total of 252 tokens were discarded for the reasons given in the table above. The remaining 308 tokens were distributed across the following DM functions, as presented in Table 10.

Table 10 - DM distribution in the revised sample

Position/DM	ALL	CNT	DCT	EVD	EXP	INP	Total
Initial	0	68	6	0	24	41	139
Medial	9	9	2	5	2	5	32
Final	30	34	2	71	0	0	137
Total	39	111	10	76	26	46	308

Considering the quantity of data and its clear imbalance across functions and positions, which might negatively affect a classification task's results, new DM candidates were searched in the texts of the matrix C-ORAL-BRASIL I corpus that was not used in the BP mini corpus. The selection criteria were prosodic isolation, position (initial or final), and lexical recurrence. For instance, proper names were identified within each file and queried for the search of potential ALL and CNT candidates. Analogously, *Nossa/No'* (holy) tokens in prosodic isolation and initial position were sought as EXP candidates. This procedure was automatically done thanks to Python scripts that read the corpus XML files and extracted the audio based on regular expressions. The resulting queries were subjected to the same revision procedures applied to Gobbo's sample. 123 new tokens were added to the sample after the new sampling. Table 11 displays the final DM distribution across the five functions. Notice that the 10 DCT tokens were excluded, since analyzing cohesive DM is out of the scope of this research.

Table 11 - Final sample used for the classification task

Position/DM	ALL	CNT	EVD	EXP	INP	Total
Initial	0	73	0	69	71	213
Medial	9	9	5	2	5	30
Final	60	57	71	0	0	188
Total	69	139	76	71	76	431

Table 12 displays the frequency of lexemes/small expressions by DM class. We see that many lexemes and small expressions can take on different DM functions.

Table 12 - Lexical frequency by DM class

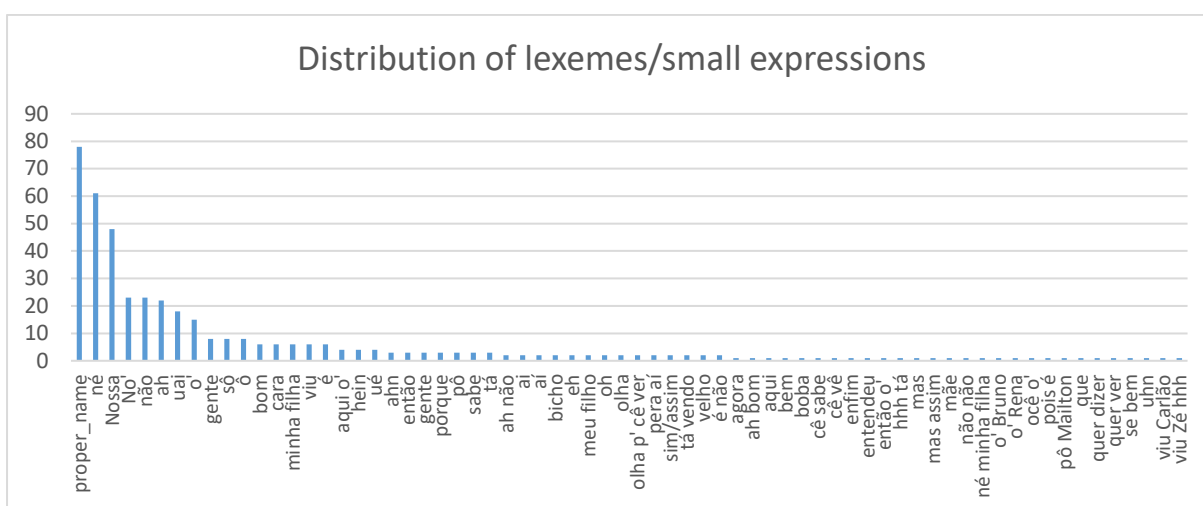
TEXT	DM FUNCTION					TOTAL
	ALL	CNT	EVD	EXP	INP	
proper_name	50	28	0	0	0	78
né	0	0	60	0	1	61
Nossa	0	1	0	39	8	48
No'	0	0	0	16	7	23
não	0	9	0	1	13	23
ah	0	6	0	8	8	22
uai	0	14	0	0	4	18
o'	0	10	0	0	5	15
gente	0	1	0	0	7	8
sô	0	8	0	0	0	8
ô	0	6	0	0	2	8
bom	0	4	0	0	2	6
cara	4	2	0	0	0	6
minha filha	6	0	0	0	0	6
viu	0	0	6	0	0	6
é	0	1	0	0	5	6
aqui o'	0	4	0	0	0	4
hein	0	0	4	0	0	4
ué	0	4	0	0	0	4
ahn	0	3	0	0	0	3
então	0	1	0	0	2	3

TEXT	DM FUNCTION					TOTAL
	ALL	CNT	EVD	EXP	INP	
gente	2	0	0	1	0	3
porque	0	1	0	0	2	3
pô	0	2	0	0	1	3
sabe	0	0	3	0	0	3
tá	0	2	1	0	0	3
ah não	0	0	0	0	2	2
ai	0	1	0	1	0	2
aí	0	1	0	0	1	2
bicho	1	1	0	0	0	2
eh	0	1	0	0	1	2
meu filho	1	1	0	0	0	2
oh	0	0	0	1	1	2
olha	0	0	0	0	2	2
olha p' cê ver	0	2	0	0	0	2
pera aí	0	1	0	0	1	2
sim/assim	0	2	0	0	0	2
tá vendo	0	2	0	0	0	2
velho	1	1	0	0	0	2
é não	0	1	0	0	1	2
agora	0	1	0	0	0	1
ah bom	0	1	0	0	0	1
aqui	0	1	0	0	0	1
bem	0	1	0	0	0	1
boba	1	0	0	0	0	1
cê sabe	0	1	0	0	0	1
cê vê	0	1	0	0	0	1
enfim	0	1	0	0	0	1
entendeu	0	0	1	0	0	1
então o'	0	1	0	0	0	1
hhh tá	0	1	0	0	0	1
mas	0	0	0	0	1	1
mas assim	0	1	0	0	0	1
mãe	1	0	0	0	0	1
não não	0	0	0	0	1	1
né minha filha	1	0	0	0	0	1
o' Bruno	0	0	0	0	1	1
o' Rena	0	1	0	0	0	1

TEXT	DM FUNCTION					TOTAL
	ALL	CNT	EVD	EXP	INP	
ocê o'	0	1	0	0	0	1
pois é	0	1	0	0	0	1
pô Mailton	0	0	0	0	1	1
que	0	1	0	0	0	1
quer dizer	0	1	0	0	0	1
quer ver	0	1	0	0	0	1
se bem	0	0	0	1	0	1
uhn	0	0	0	1	0	1
viu Carlão	0	1	0	0	0	1
viu Zé hhh	0	1	0	0	0	1
TOTAL	68	139	75	69	80	431

Figure 15 shows that the lexical fillers follow the same distribution languages generally follow: the Zipfian distribution. A few lexemes or small expressions have high frequency, and the vast majority are underrepresented. This illustrates how the lexical filling of Discourse Markers can be rather variable.

Figure 15 - Distribution of lexemes/small expressions



However, some DM classes seem to be more flexible whereas other have more constraints with respect to the lexical content. The DM classes that display the most variability in terms of lexical filling are CNT and INP. On the other hand, the most constrained class is EVD. Finally, Gobbo (2019) has shown that approximately 80% of the DMs are adjacent to their illocutionary unit, and 20% are one or more information units distant from COM.

4.3 DATA PROCESSING AND ANNOTATION

4.3.1 Data preparation

A Python script (the C-ORAL_searcher) was used to generate a list of utterances from the corpus xml files. The script extracted the portions corresponding to each utterance from the original audio files by using the start and end time information. 300 milliseconds before and after the given times were added. When audios were recorded in stereo, only the channel matching the utterance's speaker was extracted for analysis; the other channel was discarded. A Praat script created a TextGrid file (a Praat object used for segmentation and annotation) for each file using the utterance table. Five tiers made up the structure of the TextGrids (see Figure 17 - Illustration of an annotated file): (1) the transcription of the whole utterance; (2) the syllabic annotation; (3) the delimitation of the stressed vowel; (4) the transcriptions delimited by Information Unit; and (5) the tag of the Information Unit. All tiers were boundary tiers. At this point, boundaries were not yet aligned with respective events.

4.3.2 Data annotation

The annotation of syllabic units followed the criteria recommended in the literature for identifying phonetic boundaries. The oscillogram and the broadband spectrogram of the acoustic signal were simultaneously examined as a guide for the segmentation (Machač & Skarnitzl, 2009; Turk et al., 2006).

The use of ASCII characters for the phonetic transcription facilitated annotation and ensured compatibility with the normalized duration estimation method (Barbosa, 2013), detailed in the next section. A broad phonetic transcription, devoid of distinct representations for allophones, was employed, with Figure 16 illustrating the notation used and its equivalence in IPA symbols.

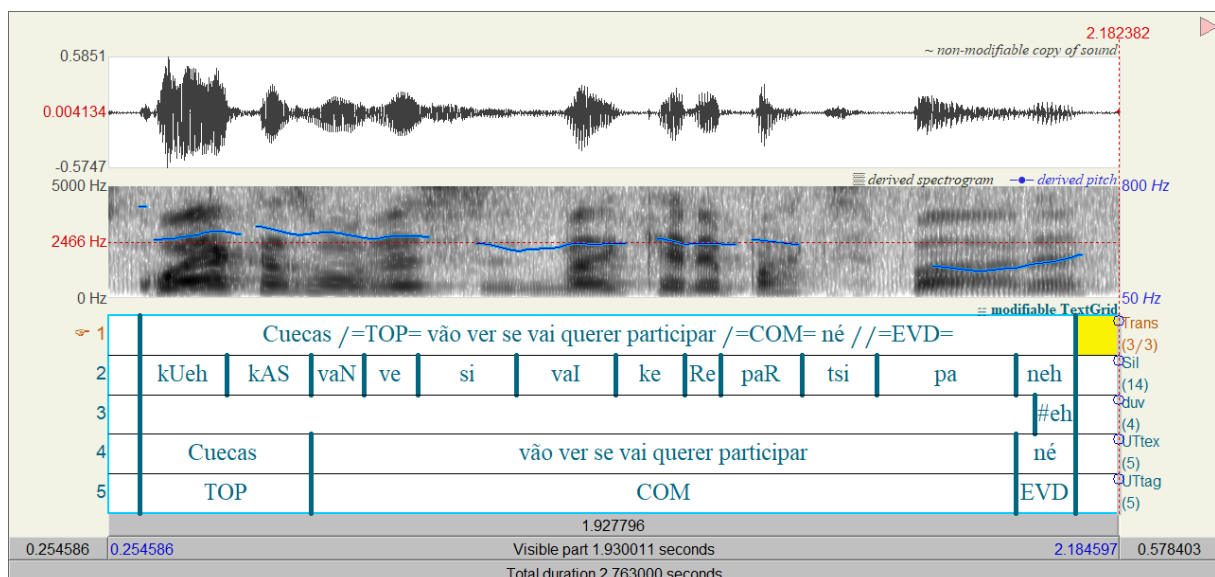
The inherent spontaneity and diverse environmental conditions of recordings generated spectrographic images more contaminated than those acquired from controlled environments with acoustic protection and muffling. This led to an increased reliance on listening to the audio files. The characteristics of utterances displaying low intensity, high articulation rate, and reduced phonetic realization exacerbate these difficulties and may result in inaccuracies.

Figure 16 - Correspondence between IPA and ASCII characters¹⁴

Correspondência IPA e Marcação ASCII						
IPA	ASCII	IPA	ASCII	IPA	ASCII	
i (tia)	i	'eɪ (sei)	eI	p (pata)	p	
e (etapa)	e	'eɪ (pastéis)	ehI	t (tapa)	t	
ɛ (série)	eh	'aɪ (pai)	aI	k (carne)	k	
a (pata)	a	'oɪ (caracóis)	ohI	b (bola)	b	
ɔ (bola)	oh	'oɪ (foi)	oI	d (dado)	d	
o (tolo)	o	'uɪ (fui)	uI	g (gosto)	g	
u (tulipa)	u	'aŋɪ (mãe)	aŋI	f (foca)	f	
'iN (inca)	iN	'oŋɪ (põe)	oŋI	ʃ (chuva)	sh	
				tʰ	th	
'eN (contente)	eN	'iʊ (riu)	iU	tʃ (ativo)	tS	
				dʒ (medida)	dZ	
'aN (cantar)	aN	'eʊ (seu)	eU	v (véu)	v	
				z (casaco)	z	
'oN (compre)	oN	'eʊ (réu)	ehU	ʒ (anjo)	zh	
'uN (junto)	uN	'aʊ (mau)	aU	s de coda (pasta)	S	
I (sete)	I	'oʊ (sol)	ohU	s de coda (mesmo)	Z	
e (ópera)	E	'oʊ (sou)	oU	m (mata)	m	
ɐ (casa)	A	'aŋʊ (não)	aŋU	n (nabo)	n	
o (cômodo)	O	Iʊ (frágil)	IU	ɲ (moinho)	nh	
u (todu)	U	uʊ (cônsul)	UU	r	r	
				ɾ (TAP)	R	
iN (intuito)	IN	ɪ̃ (série)	II	R (caipira)	Rh	
				R (glotal)	h	
				R (vibrante)	rr	
eN (hifen)	EN	ʊ̃ (tênue)	UI	l (lago)	l	
ɐN (imã)	AN	ɶ̃ (ânsia)	IA	ʎ (orgulho)	lh	
oN (conte)	ON	ʊ̃ɐ (água)	UA	ɫ	L	
uN (mundo)	UN	aŋʊ̃ (foram)	ANU			

The same challenges rendered the application of automatic segmentation procedures (forced alignment) exceedingly complex for the analyzed recordings. To be compatible with the transcription scheme used for the duration standardization procedure, the Alinha-PB phonetic aligner (Kruse & Barbosa, 2021) was tested for the task. However, satisfactory results would require more entries to the system's exception dictionary. This is because the Mineiro dialect is particularly keen on final droppings and coarticulation, anticipation, and sandhi phenomena. The Figure 17 below illustrates an annotated file:

Figure 17 - Illustration of an annotated file



The information contained in TextGrid objects was collected using the

¹⁴ For further details on the correspondence between IPA and ASCII characters, see the documentation available at: https://github.com/pabarbosa/prosody-scripts/blob/master/ProsodyDescriptorExtractor/Documents/IPAASCIICorrespondanceTable_BP.pdf

Python Praat wrapper Parselmouth package (Jadoul et al., 2018) and used to select data points of interest, which were later used to calculate the measurements described in section 4.4. Prosodic-acoustic parameters estimation).

4.4 STANDARDIZATION OF MEASURES

4.4.1 Reference for the standardization of prosodic-acoustic parameters

Prosody conveys information beyond the parsing of speech and the Information Structure. It can give cues on the attitude and speakers' emotional state or their sex, age, vocal tract characteristics, and even health conditions (Mello & Raso, 2011). This introduces variation unrelated to the linguistic functions targeted here, the Information Structure.

For this reason, it is important to adopt a procedure that inhibits non-linguistic variation as much as possible in extracting acoustic measurements. We can envisage three ways of doing this. The first would be to standardize the parameters by the mean and standard deviation of all the utterances of a given speaker. As well as making data processing computationally more expensive, this procedure may prove to be insufficient. Depending on their intentions, the speaker changes their average speech rate, intensity, and average pitch from one utterance to the next. Another procedure would be to standardize the parameters based on the averages and standard deviation of only the utterance in which the Discourse Marker is inserted. However, the IS pattern may vary in large proportions from one utterance to another. Some utterances are quite complex, and others simpler. Furthermore, in the case of the stanza units, the variation of different speech acts and adjacent structures can also be added to the basic measures for standardization.

A procedure proposed by Raso (2014) and already successfully tested in subsequent work (Raso & Vieira, 2016; Gobbo, 2019) is to take as a reference the illocutionary unit of the pattern to which the DM is linked. This procedure is advantageous because the variation in

a pattern's IS is related to this unit. The Comment (COM) is the central node that conveys the speaker's actional intention and links the adjacent informational units that supplement the speaker's communicative intention. Moreover, it is the sole IU that occurs in every terminated sequence. For future work, it could be interesting to test how the adoption of other heuristics may influence the final scores of the classification model with respect to a baseline model that takes as input non-standardized measurements. In the present work, we observed that adopting COM as a measure for standardization can help improve the accuracy of a five-class classification model by up to 30 percentage points¹⁵.

Some terminated sequences may present more than one COM unit, such as in the cases of stanzas or patterned illocutions. In the case of stanzas, we considered only the illocutionary unit in whose pattern the DM was. When the pattern had patterned illocutions, a chain of Multiple Comments (CMM), all the CMMs of the chain were considered.

4.4.2 Standardization of prosodic-acoustic parameters

For the normalization of the prosodic-acoustic parameters, Raso & Vieira (2026) and Gobbo (2019) utilized a proportional difference of the measurements from the DM with respect to the reference (COM), as shown by the formula below:

Figure 18 - Proportional difference

$$M_{DM}^d = \frac{|M_{DM}^a| - |M_{Re}^a|}{|M_{Re}^a|}$$

¹⁵ Comparing two classifications models: one trained with non-standardized features and another with standardized features.

Where:

M_{DM}^d is the value of the proportional difference between the DM and the Reference Unit;

M_{DM}^a is the absolute value of the DM; and

M_{Re}^a is the absolute value of the Reference Unit.

This procedure was applied to the consolidated values (statistics) of the entire DM and Reference Units. For each audio file (containing one utterance), I estimated the raw prosodic measurements with a 5ms sampling rate. Then, based on the annotation files and the temporal information of the target DMs and reference units' boundaries, I set off time points of interest and obtained consolidated measurements. For this reason, I applied the normalization procedure frame-wise, before consolidating the measurements. Unlike previous work (Gobbo, 2019), Standard Score (Z-scores) normalization procedure was adopted. Each sampled time point was transformed into z-scores considering each utterance's mean and standard deviation previously computed for the Reference Units (COM/COB/CMM annotated as such in the TextGrids). The z-score is obtained by the following formula:

$$f(t) = \frac{x_t - \mu_{Re}}{\sigma_{Re}}$$

Where:

$f(t)$ is the function for transforming each time point in z-scores;

x_t is the estimated measure of each time point;

μ_{Re} is the mean of the estimated measure for the Reference Unit's

interval; and

σ_{Re} is the standard deviation of the estimated measure for the Reference Unit's interval.

This procedure was applied to all estimates before deriving the prosodic descriptors of intensity, duration, and fundamental frequency outlined in the next section.

4.5 PROSODIC-ACOUSTIC PARAMETERS ESTIMATION

The estimation of prosodic features outlined in this section mostly followed the ones proposed in Gobbo (2019). I added parameters of f0 curves and some parameters of intensity (spectral emphasis and intensity in voiced regions) and made some changes concerning how features were calculated. In total, 30 features were derived from the standardized measurements. They are organized as follows:

- a) features of intensity;
- b) features of duration;
- c) features of f0;
- d) features of f0 variation;
- e) features of alignment with the DM's stressed vowel;
- f) features of f0 curve.

The extraction of estimations for each time point was automated through Python and Praat scripts and then processed through a Python script for target intervals.

4.5.1 Features of intensity

Intensity features were extracted using a Praat wrapper Python package from intensity objects. We derived six features of intensity:

- a) Mean intensity of the DM;
- b) Intensity standard deviation of the DM;
- c) Maximum intensity of the DM;
- d) Minimum intensity of the DM;
- e) Mean intensity on the DM's stressed vowel¹⁶;
- f) Mean spectral emphasis on the DM's stressed vowel¹⁷;

Spectral emphasis was calculated as in Traunmüller & Eriksson (2000), considering values for the stressed vowel.

4.5.2 Features of duration

Duration features were:

- a) Mean Z-scores of the DM's syllabic duration;
- b) Z-scores of the DM's stressed syllable;
- c) DM's raw duration;

¹⁶ Considering only voiced regions of the audio file.

¹⁷ Considering only voiced regions of the audio file.

The estimation of z-scores for the phonetic syllables followed Barbosa (2013) using a Praat script implemented by Gobbo (2019). The data was then consolidated with a Python script. We did not consider the standardized duration of the DM unit proposed by Gobbo (2019) since it correlates with the mean standardized duration and since that feature is irrelevant in Gobbo’s three-class model.

Modeling the relationship between perceived rhythm and articulation rate is also non-trivial. This is due to phoneme-specific intrinsic and co-intrinsic durations. Several procedures are available to normalize the raw duration of segments in similar contexts (Campbell & Isard, 1991). This statistical approach allows an efficient estimation of segmental lengthening, expressed as the deviation from the expected duration of a phoneme with a set of properties. Relying on this approach, Barbosa (2007) developed a model of speech rhythm that estimates the lengthening of syllable or syllable-like units (the so-called Vowel-to-Vowel unit – or VV unit) at the segmental level.

This algorithm considers in-context expected mean durations and standard deviations and serially applies two techniques for normalizing raw duration: a z-score transformation and a 5-point moving average filtering procedure. The z-scores are calculated according to the equation 2 below:

Equation 1 – Z-scores

$$z = \frac{(x - \mu)}{\sigma}$$

where z , the z-score for a given segment, is calculated by subtracting μ , the mean expected value of this segment, from x , the raw duration value, then dividing by σ , the standard deviation of the expected value. Each z_i z-score value is then smoothed by the moving average filter. The closer to z_i , the larger the weight applied to neighboring z-scores

(eq. 3).

Equation 2 – Z-smoothen-i

$$z_i = \frac{1 \cdot z_{i-2} + 3 \cdot z_{i-1} + 5 \cdot z_{i+3} + 3 \cdot z_{i+1} + 1 \cdot z_{i+2}}{13}$$

This model was implemented into a semi-automatic tool available to the research community, that outputs measures of rhythm from segmented speech (Barbosa, 2013).

4.5.3 Features of fundamental frequency (f0)

The processing of f0 data points was the object of a special procedure described in a dedicated chapter. All measurements involving f0 data points were calculated from pre-processed f0 estimations. The purpose of this procedure was to avoid manual intervention for the f0 estimation and tracking, as was done in Gobbo (2019). The features of f0 are:

- a) Mean fundamental frequency (f0) of the DM;
- b) Standard deviation of f0 of the DM;
- c) Maximum f0 estimation of the DM;
- d) Minimum f0 estimation of the DM;

4.5.4 Features of f0 variation

The measures of f0 variation were:

- a) F0 slope over the DM from the beginning to the ending points;
- b) F0 slope over the DM's stressed vowel;
- c) F0 range over the DM;
- d) F0 slope before the central point of the stressed vowel;
- e) F0 slope after the central point of the stressed vowel;

F0 slope measures were calculated using the linear coefficient outputted by the polyfit function of the Numpy Python package (Harris et al., 2020). F0 points were separated by region of interest: the complete DM (a); the stressed vowel (b); the values before the central point of the stressed vowel (whose boundaries were annotated in a special tier on Praat); the values after the central point of the stressed vowel. The F0 range within the DM was calculated as the difference between the maximum and minimum f0 estimations over the DM.

A difference with respect to Gobbo (2019) is that the slopes were not calculated by taking the f0 estimates for initial and ending times but by fitting the regression line to all the points available for the respective intervals. This has the advantage that potentially deviant points at the boundaries of regions of interest will not have a relevant impact on the computation (note that micro-prosodic effects are likely to be observed at segmental boundaries).

4.5.5 Alignment features

The alignment features were:

- a) Ratio of maximum intensity;
- b) Ratio of minimum intensity;
- c) Ratio of maximum f0;

- d) Ratio of minimum f_0 ;
- e) Ratio of maximum intensity with respect to the central point of the stressed vowel;
- f) Ratio of minimum intensity with respect to the central point of the stressed vowel;
- g) Ratio of maximum f_0 with respect to the central point of the stressed vowel;
- h) Ratio of minimum f_0 with respect to the central point of the stressed vowel.

From (a) to (d), the ratios were calculated as the difference between the timing of the critical point (max or min of f_0 and intensity) and the timing of the initial point of the DM, divided by the duration of the DM. These features aimed to show the proportion, from 0 to 1, at which these critical points were realized.

From (e) to (h), the ratios were calculated as the difference between the critical point's time and the vowel's starting point, divided by the duration of the stressed vowel. Unlike the previous measurements, these measurements can take on values inferior to 0 or superior to 1. Values between 0 and 1 indicate that the critical point occurred within the stressed vowel. If they equal 0, the critical point happens at the beginning of the stressed vowel, and if they equal 1, exactly at the end of the stressed vowel. Values inferior to 0 indicate critical points before the stressed vowel, and values superior to 1 refer to critical points after the stressed vowel.

4.5.6 Features of f_0 curves

A vector of 30 f_0 data points was estimated for each DM using linear interpolation (through the `interp1D` function in Python) to run a curve-

fitting algorithm with a normalized temporal vector (see Xu, 2013). The procedure to obtain the curve coefficients is further described in section 7.2. Curve fitting). The number of 30 points was chosen by dividing 300ms (approximately the average duration of a Discourse Marker) in 10ms steps. Further details on the curve fitting are given in section 7.2.

5 EXTRACTING ROBUST F0 CURVES

5.1 MOTIVATION

Estimating the fundamental frequency (f_0) has historically been challenging in audio signal processing. While numerous context- and condition-specific approaches have been developed and have been successful in their particular uses, creating context- and condition-free f_0 estimators is a rather bold task (Gerhard, 2003; Raso et al., 2022). For example, an all-purpose PDA (musical note and speech detection) without Viterbi smoothing, like YIN (de Cheveigné & Kawahara, 2002), may perform less efficiently when applied for speech analysis; a noise-resilient PDA (BaNa - Ba et al., 2012; Yang et al., 2014) may fail to properly devoice zones of the audio affected by reverberation phenomena, since this algorithm tends to overestimate the number of voiced frames in clean conditions. As a result, while there are many f_0 estimators available today, very few (if any) of them will be able to cover and solve, at the same time, all or the majority of the problems affecting the f_0 signal.

The correct estimation and tracking¹⁸ of f_0 values become an even greater problem when we deal with spontaneous speech data. Recorded out of acoustically isolated and controlled settings, the signal of spontaneous speech can be rapidly degraded by non-modal phonation (Gerratt & Kreiman, 2001) and the additive noise it may generate (D’Alessandro, 2006), other analog phenomena, and for technical reasons¹⁹. Non-modal phonations raise significant difficulties regarding evaluating a frequency (i.e., a regular phenomenon), whose

¹⁸ By *f_0 estimation*, we mean the estimation of alternative f_0 values within a unique timeframe (each f_0 candidate value); *tracking* stands for the extraction of continuous f_0 trajectories (considering all timeframes) from the underlying sources (possible f_0 candidates at each timeframe).

¹⁹ Saturation, low sampling rate and low bandwidth, lossy audio compression.

definition is problematic when vocal folds vibration mechanisms are not regular.

Several methods and algorithms have been proposed to estimate formant values and to determine, at each timeframe, what the most probable f_0 candidates are. Each of these PDAs is sensitive to different phenomena degrading the speech signal. For instance, some of them can identify the best candidates even in very noisy conditions, while others may not return any estimation. Of course, solving all these problems is a fiendishly difficult task, which, by no means, we imply to undertake. All the same, we do want to put to good use the strengths of available PDA algorithms to have better f_0 estimations and tracking in a natural setting.

Another problem is that speech data recorded in non-controlled settings requires the researcher to tweak many parameters offered by PDA algorithms. Since each recording setting displays different acoustic conditions, each audio might need different parametrization to produce the most realistic f_0 estimation and tracking. By way of example, linguists who have dealt with spontaneous speech can easily relate to the fact that each audio may need a different voicing threshold value to display the most adequate voicing decision as perceived for the segmental material. This solution will likely replace the file-by-file parametrization required to obtain appropriate f_0 estimations, tracking, and adequate voicing decision.

5.2 PITCH DETECTION ALGORITHMS

Most PDA algorithms can be classified into four broad categories: (a) time-domain methods, which are based on the temporal dynamics of the signal; (b) frequency-domain methods; (c), hybrid methods, which put together time- and frequency-domain approaches; and (d) statistical frequency-domain methods. Except for statistical methods (for instance, approaches based on neural networks), these approaches have in common the fact that they (a) pre-process the signal by filtering or splitting it into frames, (b) search for values most likely to be f_0 candidates, and (c) track the most probable f_0 trajectory or

impose transitional constraints so as to output estimations and tracking that display a continuity – since in each timeframe we have several competing f_0 candidates whose strengths will not necessarily be in a continuous curve.

In this section, we give a rather elementary overview of how some of these algorithms work. Several premises and mathematical explanations are deliberately skipped since they would entail a much more profound and complex research. For more in-depth and comprehensive accounts of the main algorithms, implementations, criticisms, and performances, we refer the reader to Gerhard (2003), Ferro & Tamburini (2019), Sukhostat & Imamverdiyev (2015), Jouvét & Laprie (2017), and Bechtold (2021), as well as to each PDA's main references, which provide, besides implementation detailing, benchmarking on competing PDAs' performances.

One of the simplest approaches to estimating f_0 is that of zero-crossing algorithms. It consists basically in measuring the timing between the signal's zero-crossing points and calculate the frequency by dividing the number of complete periods per unit of time (complete periods per second). But this approach is not reliable when the signal displays complicated waveforms which are composed of multiple sine waves with differing periods, and, especially in noisy data – which is the case of speech data.

More complex methods search for matches by comparing portions of the signal with other portions that have been offset by a trial period. This is how auto-correlation algorithms like AMDF (average magnitude difference function), AS MDF (average squared mean difference function), and others functions work. For signals with high periodicity, these algorithms can produce estimations that are quite accurate. However, the most basic implementations struggle with noisy data and are often prone to octave and fifth jumps. Time-domain approaches frequently build upon these methods. But to overcome the above-mentioned difficulties, they come with smoothing procedures and movement constraints to make their estimations and tracking more in line with how humans would judge the pitch.

On the other hand, the periodogram is used by frequency-

domain methods to translate the signal into a rough estimate of the frequency spectrum. The Fast-Fourier Transform (FFT), the crucial component of the periodogram method, makes such approaches suitable for many tasks. Nonetheless, the algorithms build upon this approach demands more processing capacity as the necessary precision levels grow. Well-known frequency-domain algorithms include the harmonic product spectrum, cepstral analysis, maximum likelihood, and the detection of peaks produced by harmonic series.

As mentioned, some algorithms put together temporal and frequency-based approaches. These algorithms are based upon a combination of time-domain processing using an autocorrelation function like the normalized cross-correlation function and frequency-domain processing based on spectral information in order to identify f_0 candidates. Having f_0 candidates from both time and frequency domains, the algorithm may use a dynamic programming algorithm or other movement constraints to output a final f_0 tracking.

Recently, PDA algorithms that use statistical approaches to do part of the estimation or the tracking tasks have been put forward. By way of example, Ferro & Tamburini (2019) proposes a Neural Smoother intended to improve the performances of other PDA algorithms. This smoother is intended to be an additional layer of postprocessing, as the author acknowledges that most PDA algorithms are accompanied by some Viterbi-like smoothing procedure or movement constraint after f_0 estimation. The authors point that postprocessing procedures are often insufficient to produce a reliable f_0 contour throughout whole utterances. The proposed postprocessing smoother acts by leveraging a Long Short-Term Memories Neural Network, a particular kind of recurrent neural network, used to correct f_0 detection errors outputted by state-of-the-art Pitch Detection Algorithms. One criticism to such approach is that, even if a good-performance model is found, it does not allow for the understanding of how the problem is solved (black-box model). As pointed by Gerhardt (2003), the algorithmic information ends up stored in the model's weights/parameters, hindering the comprehension of what is going on under the hood. Especially in models based on deep, complex architectures, the mapping between the variables passed on to the

model and its final weights becomes rather opaque.

As an initial step, an inclusive approach is adopted, where multiple Pitch Detection Algorithm (PDA) algorithms, spanning time-, frequency-, and mixed domains, are employed. This approach, described as a hit-on-everything-that-walks strategy, aims to generate f_0 estimations that are more independent of context and conditions. The outputs from these diverse algorithms are then compared to determine the most plausible estimation. Utilizing a dynamic programming approach, the best f_0 tracking is established, incorporating additional penalizing and rewarding factors. Subsequently, a Neural Network (NN) model is applied to derive a context-independent, frame-by-frame voicing decision.

Given the significant computational cost associated with employing numerous algorithms, a crucial aspect of this work involves eliminating redundant information, particularly from more intricate algorithms. In the forthcoming sections, the algorithms employed thus far are enumerated, and a comprehensive explanation of the Voicing Decision model is provided.

5.3 VOICING DECISION MODEL

5.3.1 Introduction

Studying prosodic changes in spontaneous speech is increasingly important – especially when the target concerns pragmatic variations linked to sociolinguistic factors, real-life interactions, and spoken expressivity (see, e.g., Drager, 2015; Émond et al., 2013; Meer & Fuchs, 2022). A main problem with spontaneous speech is linked to the frequently poor recording quality, as well as the presence of reverberation and various background noises (e.g., street noise, ambient noise in busy places, background voices or music).

Obtaining reliable f_0 profiles is a task that can be split into three

subtasks: f0 values estimation, f0 tracking, and Voicing Decision (VD)²⁰. As previously said, to accomplish these tasks, many pitch detection algorithms (PDA) are available today with good performances. However, most PDAs do not behave well in noisy situations. A rapid degradation of performances may be observed when different kinds of noise are added to clean speech (Émond et al., 2013; Jouvét & Laprie, 2017). Jouvét & Laprie (2017) also noticed that the type of error (estimation or voicing decision) varies according to the type of noise.

Some PDAs were specifically developed to be robust to noise (e.g., Gonzalez & Brookes, 2014; Yang et al., 2014). However, some still fail or were not designed to tackle the Voicing Decision task specifically. For instance, Yang et al. (2014) presents a good performance for f0 estimation in noise, but the voicing decision (VD) task is left aside.

As pointed out by Jouvét & Laprie (2017), the VD task is one of the main sources of errors when PDAs are assessed in noisy conditions. This task has recently received some attention as a stand-alone object (see e.g., Batra et al., 2022, for clean audio, and Pradeep et al., 2019, for white noise). This section shows that the VD accuracy of available PDAs is differently affected by the type of noise added. Some PDAs perform better on clean data but do not yield good results on noisy data; others were specially developed to be robust to noisy data but underperform on clean data. To solve this issue, a CNN-based classification model was trained. Its aim is to outperform the VD accuracy of tested PDAs on both clean and noisy data. The specificity of the system is that it was trained on a much more varied range of types and levels of realistic noises than those of previous studies.

Some phenomena differ in controlled and natural conditions (see, e.g., de Ruyter, 2015; Meer & Fuchs, 2022) and spontaneous speech (broadcast media or field recordings). These phenomena need to be well understood in order to adequately process spoken communication in various settings (Wagner et al., 2015). For this reason,

²⁰ Note that the voicing or voiced/unvoiced decision here is not the same task as Voice Activity Detection (VAD) (e.g., Lavechin et al., 2020) that targets longer-term detection of speech, may it be voiced or not.

the VD accuracy of the assessed PDAs and the proposed system on a corpus of spontaneous speech recorded in natural settings are also evaluated.

The methods section presents (a) the corpora used and how they were augmented with different types and levels of noises, (b) the set of features used to train the proposed model, and (c) the models' architecture. In the results section, the evolution of VD errors for a set of existing PDAs and the proposed model are evaluated to compare their performances in various types of noise and signal-to-noise ratio (SNR) levels and on real spontaneous (and noisy) datasets.

5.3.2 Methods

5.3.2.1 Corpora

Three speech databases were used (KeelePitchDB, CSTR, and C-ORAL-BRASIL-I); two of them provide clean speech with electroglottogram signal to allow for accurate F0 measurements, and one is a sample of a spontaneous speech corpus that was only used for evaluation. The C-ORAL-BRASIL corpus (Raso & Mello, 2012) is a large database of spontaneous speech containing recordings of dozens of Brazilian Portuguese speakers in various styles, situations, and places. Details on the corpora can be found in Table 1. The two clean corpora were augmented by adding various types of noise to the original recordings at different signal-to-noise ratios (SNR). From the C-ORAL complete corpus, a sample was selected for a study on Discourse Markers. From there, 62 audio files that feature many natural background noises were picked to serve as unseen data collected in natural settings. To establish their voicing ground truth, audio files were manually revised at frame level by two experienced annotators whose inter-annotator agreement degree was assessed as almost perfect - Cohen's Kappa (Cohen, 1968): 0.8; agreement rate: 0.91.

Table 13 - Speech corpora: name, language (Lang: English or Brazilian

Portuguese, BP), number of speakers (Spk: Fe- male/Male), total duration (Dur, in minutes), proportion used for training and testing (Tr/Te, if applicable), and reference (Ref)

Name	Lang	Spk	Dur	Tr/Te	Ref
KeelePitchDB	English	5/5	5	80/20	(Plante et al., 1995)
CSTR	English	1/1	5	80/20	(Bagshaw et al., 1993)
C-ORAL-BRASIL	BP	11/14	3.7	-/100	(Raso & Mello, 2012)

The two clean corpora were augmented by adding various types of noise at different signal-to-noise ratios (SNR) to the original recordings. From the C-ORAL complete corpus, a sample was selected for a study on Discourse Markers (Gobbo, 2019). From there, 62 audio files that feature a large amount of natural background noises were picked to serve as unseen data collected in natural settings.

The dataset used for training and testing was produced by augmenting the two clean corpora (KeelePitchDB and CSTR in Table 13). Eight different noises were applied to each original sound (extracted from the RSG-10²¹ and QUT-NOISE-TIMIT²²) and two room-impulse-answers (extracted from the C4DM Room Impulse Answer²³ database) to introduce reverberation at different SNRs. These particular noise types were chosen for two reasons: (a) they frequently occur in recordings made in a natural setting, and (b) they degrade f0 estimation and VD on different levels. Details on the noises are given in

Table 14. The nine targeted SNRs were 20, 15, 10, 5, 0, -5, -10, -15, -20dB. Before the processing, all sounds were down sampled to

²¹ <http://www.steeneken.nl/7-noise-data-base/>

²² <https://github.com/qutsaivt/QUT-NOISE>

²³ <http://www.isophonics.org/content/room-impulse-answer-data-set>

16kHz.

Table 14 - Types of noise and source

Name	Type	Origin	Dur	Ref
Classroom	RIR	C4DM	–	(Stewart & Sandler, 2010)
Large room	RIR	C4DM	–	(Stewart & Sandler, 2010)
Babbling (F)	noise	RSG-10	3'1"	(Steeneken & Varga, 1993; Varga & Steeneken, 1993)
Babbling (M)	noise	RSG-10	3'55"	(Steeneken & Varga, 1993; Varga & Steeneken, 1993)
CAFE-CAFE-1	noise	QUT	42'	(Dean et al., 2010, 2015)
CAR-WINDOWNB-1	noise	QUT	44'	(Dean et al., 2010, 2015)
STREET-CITY-1	noise	QUT	32'	(Dean et al., 2010, 2015)
Air Cond.	noise	Lab	4'	*
Ventilator	noise	Lab	4'	*
Electr.	noise	Lab	4'	*

The noise augmentation was performed using Praat scripts (Boersma & Weenink, 2022). Each original sound was adjusted to a default, arbitrary mean (over the complete file) intensity level of 70dB (approx. -21 dBFS), and each noise to a mean intensity level corresponding to one of the targeted SNR (i.e., at a level of 65 dB to reach a mean SNR of 5 dB). The two sounds were then mixed. For each original sound, an extract of equal duration as the target sound was extracted from a random part of the noise file and used for the nine SNR values of this given noise. For the reverberation, a convolution between the original sound and the RIR was performed; next, the produced signal was mixed with the original sound using a similar process as for the noise, at a given SNR - the mean level of the convoluted sound being adjusted to obtain the desired SNR (the extra length was cut before mixing it). In this process, the original dataset was augmented by a factor of 90.

From this augmented dataset, about 20% was pseudo-randomly selected for testing. Two speakers (one female and one male

out of 10 speakers) were randomly selected from the KeelPitchDB (which contains only one file per speaker), and 20% of the files of each of the two speakers from the CSTR database. Then, these original files, and all those obtained through the noise augmentation process, were grouped to be used as a test set, while the remaining were used for training. This way, 20% of the speakers of the KeelPitchDB were not seen by the models, and neither were the sentences of CSTR (having only two speakers in this last corpus, it is impossible to separate one from the training set). The C-ORAL-BRASIL subset, recorded in noisy conditions and thus not augmented, was used as unseen data for the final evaluation.

5.3.2.2 *F0 estimation and voicing decision*

The estimation of f_0 benefits from a large inventory of existing PDAs. Most PDAs can be classified into three broad categories with respect to f_0 estimation: (a) time-domain methods (TD), which are based on the temporal dynamics of the signal; (b) frequency-domain methods (FD); and (c) hybrid methods, which put together time- and frequency-domain approaches. Besides, PDAs may deploy techniques to improve f_0 tracking (such as smoothing and the Viterbi algorithm - VA) and reach a voicing decision. A set of 14 PDAs were chosen for their availability and the variety of their approach to f_0 estimation and voicing decisions. Table 15 details the PDA used in this study. Each PDA was used to estimate the f_0 from the waveforms of the augmented corpus. F_0 outputs may greatly vary depending on the f_0 range passed on to PDA algorithms.

Table 15 - List of the PDA tested in this study, with general characteristics

PDA	Type	References
Praat AC	TD	(Boersma & Weenink, 2022)
Praat SHS	FD	(Boersma & Weenink, 2022; Hermes, 1988)
Praat CC	TD+VA	(Boersma & Weenink, 2022)

PDA	Type	References
RAPT	TD+VA	(Talkin et al., 1995)
YIN	TD	(de Cheveigné & Kawahara, 2002)
Legacy STRAIGHT	FD+TD	(Kawahara et al., 2005)
SWIPE	FD	(Camacho & Harris, 2008)
SWIPEP	FD	(Camacho & Harris, 2008)
YAAPT	FD+TD+VA	(Zahorian & Hu, 2008)
openSMILE AC	TD+VA	(Eyben et al., 2013)
BaNa	FD+TD+V	(Yang et al., 2014)
PEFAC	FD+VA	(Gonzalez & Brookes, 2014)
pYIN	TD+VA	(Mauch & Dixon, 2014)
SRH (COVAREP)	FD	(Degottex et al., 2014)

Although some approaches rely on default ranges (see, e.g., Vaysse et al., 2022), we selected a broad range (75 to 1000 Hz)²⁴. Given the nature of the data, f0 peaks as high as 850Hz were observed in the C-ORAL-BRASIL subset. Other default parameters were not changed. Among the tested PDAs, some algorithms produce f0 estimations for each frame, plus a voicing probability, while others only output f0 estimates for voiced frames (i.e., after explicitly having a voiced/unvoiced decision). When a voicing probability was available, different probability thresholds were tested to maximize their accuracy score with respect to the ground truth. To that effect, I used a search space of seven probability levels distributed between the second and third observed probability quartiles.

5.3.2.3 Additional features

Acoustic features were extracted from the speech signals to train the VD model. The first set was the f0 values (and VD decision) estimated

²⁴ The simpler implementation of de Cheveigné & Kawahara (2002) used here, does not allow for a maximum range value.

by the 14 PDA systems. We also estimated the intensity and the Cepstral Peak Prominence using Praat (Boersma & Weenink, 2022), the Harmonics-to-Noise Ratio using the PAPD algorithm (Sturmel, 2011), the spectral emphasis following (Traunmüller & Eriksson, 2000), and 20 MFCCs using the librosa package (McFee et al., 2015) for Python. Outputs of all features were resampled at a 10ms step.

5.3.3 Models

Prior to adopting a specific modeling technique, we tested several baseline models by using a subset of our data on an 80/20 split. The results (accuracy and f1 scores) are shown in Table 16:

Table 16 - Performance of tested models

Model	Accuracy	F1-score
Stochastic Gradient Descent	69	69
LogReg	69	73
LDA	69	73
GaussianNB	70	73
RandomForest	72	75
RNN LSTM	84	86
FNN	85	87
RNN GRU	85	87
RNN BiLSTM	85	87
CNN	85	88

The obtained accuracies led us toward the use of a CNN model. Although RNN models exhibit similar results, the CNN model has two advantages: (a) it can account for neighboring timeframes (like the RNN models), which is desirable since providing context may improve the model (Hinton et al., 2012), and (b) CNN architectures offer similar results with more efficiency. Furthermore, it yielded higher accuracy and f1-score. The CNN model is configured as follows:

(A) Conv2D: Filters: 32; Kernel size: 3x3; ReLU.

(B) Conv2D: Filters: 64; Kernel size: 3x3; ReLU.

(C) MaxPooling2D: Pool size: 2x2; followed by Dropout 01 and a Flatten layer.

(D) Dense: Units(s): 32; ReLU; followed by Dropout 02.

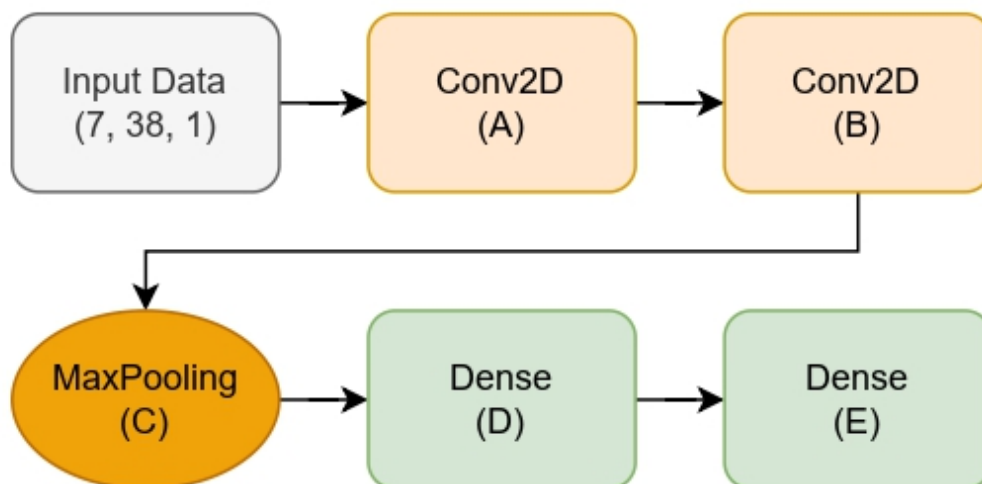
(E) Dense: Unit(s): 1; Sigmoid activation function.

A Label Smoothing is applied to the output.

An Early stopping callback was set to monitor validation loss value with patience of 5 increasing values.

Figure 19 displays the corresponding layers of the model:

Figure 19 - Flowchart of the VD CNN model architecture



To find the best parameters for the model, a Hyperband search (Li et al., 2017) was used. The search spaces, as well as the chosen parameters, are shown in the last column of Table 17.

Table 17 - Hyperparameters, search spaces and selected parameter for the VD decision CNN model

Hyperparameter	Search Space	Sel. par.
Batch size	16, 32, 64, 128, 256	64
Optimizer	Adam, RMSprop, Adadelta	Adam
Optimizer LR	0.0001, 0.001, 0.01	0.0001
Droupout 01	0.0, 0.2, 0.4, 0.6, 0.8	0.2
Droupout 02	0.0, 0.2, 0.4, 0.6, 0.8	0.5
Label Smoothing	0.0, 0.1, 0.2	0.1

A 5-fold cross-validation was conducted to evaluate the performance of our best model by splitting the dataset into five equal parts. For each iteration, the model was trained on four folds and evaluated on the remaining fold. The process was repeated for each different held-out fold for evaluation. To avoid data leaking, speakers and files were not shared between train/test splits. Our model achieved an average accuracy of 88.21% over the five iterations, with a standard deviation of 0.61. This result indicates that the model is consistent in its performance across different parts of the dataset.

5.3.4 Results

This section presents two sets of results: one linked to the degradations of the VD (by PDAs and the selected models) on controlled additions of noises, evaluated on the test set of the augmented corpus, and another is their output on the C-ORAL-BRASIL subset, which offers a naturally occurring set of noisy speech. BaNa (Yang et al., 2014), YIN (de Cheveigné & Kawahara, 2002), and RAPT (Talkin, 2005) are excluded from the comparison since they target f0 estimation but not VD.

The quality measurement used here is the Accuracy calculated as the ratio of true voiced and unvoiced frames to the total number of frames; this is the inverse of the Voicing Detection Error used by (Jouvet & Laprie, 2017). Results are shown in Table 18:

Table 18 - Global Accuracy (Glob. acc.) observed on the test set for each PDA and Model: mean (standard deviation), all SNR and noise type mixed; Accuracy of these systems on the Clean part of the test set only (Clean Ac.); Accuracy estimated on the Unseen data

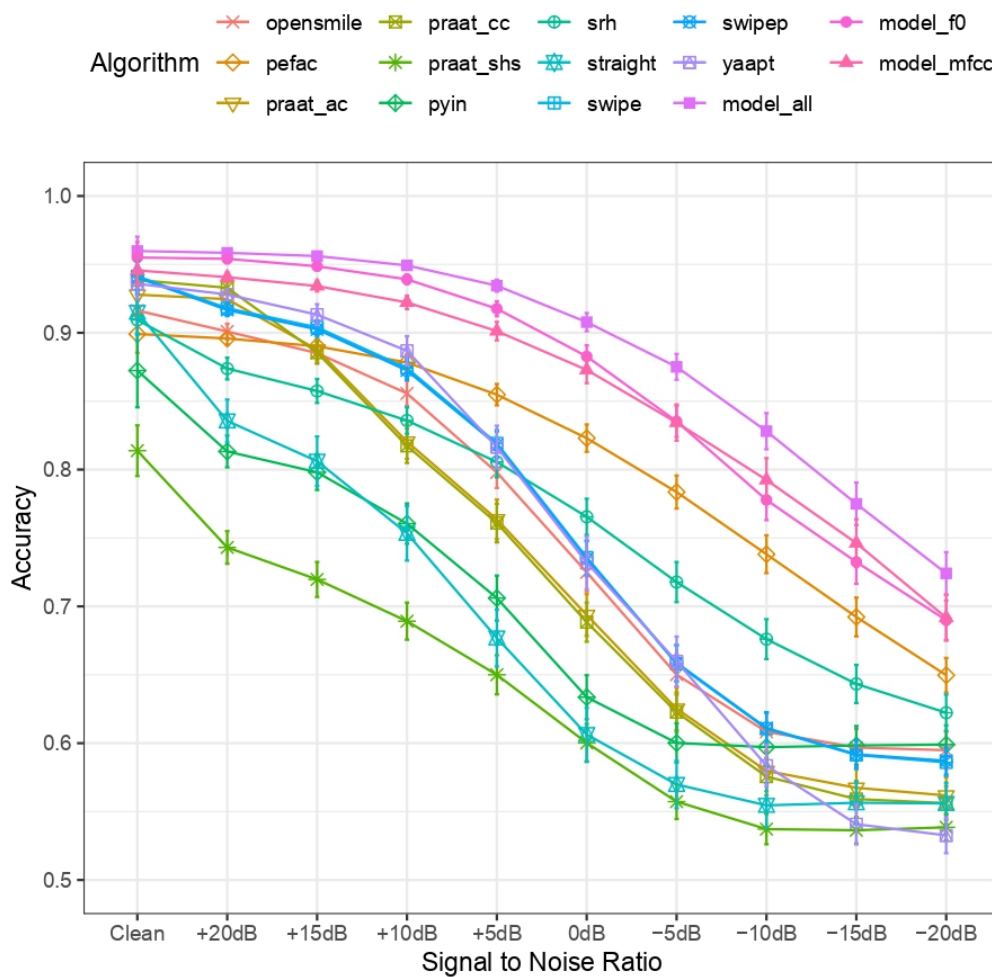
PDA/Model	Global accuracy		Clean accuracy		Unseen accuracy	
	Mean	Std.	Mean	Std.	Mean	Std.
Praat AC	0.72	(0.17)	0.93	(0.03)	0.84	(0.10)
Praat CC	0.71	(0.17)	0.94	(0.02)	0.83	(0.10)
Praat SHS	0.62	(0.12)	0.81	(0.04)	0.69	(0.12)
Straight	0.66	(0.17)	0.91	(0.04)	0.74	(0.16)
Swipe	0.75	(0.15)	0.94	(0.03)	0.83	(0.09)
Swipep	0.75	(0.15)	0.94	(0.03)	0.83	(0.08)
YAAPT	0.73	(0.18)	0.94	(0.03)	0.85	(0.08)
openSMILE	0.74	(0.15)	0.92	(0.04)	0.76	(0.16)
PEFAC	0.80	(0.11)	0.90	(0.03)	0.86	(0.05)
pYIN	0.68	(0.14)	0.87	(0.06)	0.66	(0.19)
SRH	0.76	(0.13)	0.91	(0.03)	0.73	(0.22)
Model all	0.88	(0.11)	0.96	(0.02)	0.88	(0.06)
Model f0	0.85	(0.12)	0.96	(0.03)	0.88	(0.05)
Model MFCC	0.85	(0.12)	0.95	(0.02)	0.77	(0.12)

5.3.4.1 Effect of noises at different SNR

Figure 20 presents the expected degradation of performances, in terms of Accuracy, observed on the test set with increased SNR for all PDAs and for the three models --- according to the type of noises considered. Most PDAs are robust to light levels of noise, but

performances generally drop between SNR +15 and +10 dB. The Praat AC system, which has among the best performances on clean speech (see the Clean Acc. column of Table 18), is relatively sensitive to noise: its performances rapidly degrade on noisy signals (at about +10dB SNR). Conversely, the PEFAC model has remarkably robust performances in noise (being the best PDA tested here at SNRs below 10dB), but has lower performances on clean speech. These different performances of available PDAs support our approach to build a system that could be accurate on any signal, clean or noisy.

Figure 20 - Effect of SNR (all noises mixed) on the accuracy of the VD by the 11 PDAs and the three models.



Comparing the curves of the three models in Figure 20, one can observe the complete model outperforms the others --- with mean accuracy above 0.9 at SNR = 0dB, which is a remarkable performance: there is, thus, some synergy between the proposed features. The model based on f0 only comes close to the full one in "clean" situations (because the PDAs already did a great job), while the MFCC features have more importance in the most adverse situations.

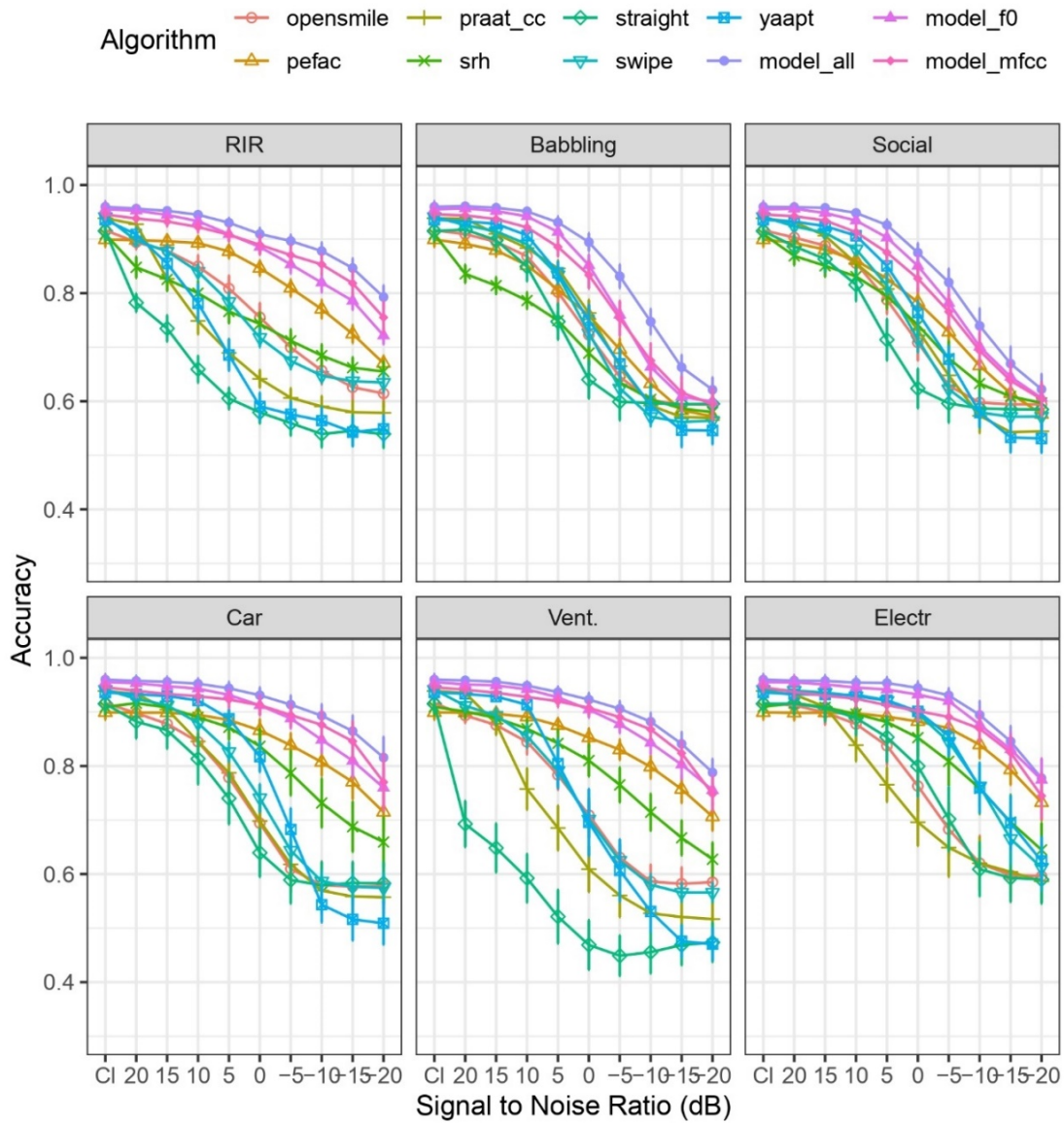
5.3.4.2 *Effect of noise type*

Figure 21 presents the slope of Accuracy with SNR for a sub-selection of PDAs (among those best performing globally for this task) and for the three models, according to the type of noises considered.

Noises have been grouped according to their characteristics: RIR includes the two reverberations, babbling the two babble noises, Social regroups Café and Street (that have non-stationary characteristics), and Vent. includes Ventilator and Air conditioning. Comparing the results on the different noise types in Figure 21, one can observe large differences: it is important to propose systems trained in noise conditions adequate to real recording situations. While reverberation, which is an adverse condition for most PDAs, is well-supported by our models, the two more difficult situations seem to be related to babbling and social noises (let's note the best model is still close to 0.9 at SNR = 0dB – but performances fall afterward).

Figure 21 - Accuracy per SNR level and noise type (plots), for our three

models (all, f0, MFCC) and the tested PDAs



The other noises (on the lower row of Figure 21, which have some stationary characteristics, but may introduce periodic noise, are also detrimental to many PDA, even at the lightest noise levels --- but they are also well dealt for by our proposed models.

5.3.4.3 *Evaluation with the spontaneous speech corpus*

The evaluation of unseen data was performed on a spontaneous corpus that features some files with notable noise. The column Unseen Accuracy of Table 18 gives the mean performances of the PDA and of the models on this dataset. Compared with performances on clean data, it shows a significant degradation of performances for all systems. This comparative degradation from a clean dataset to a noisy one shows the proposed model still provides the best performances, and a relatively reduced degradation (-8 p.p.), but the PEFAC system shows its strength in such a situation, having comparable scores, and the smallest degradation (having relatively low scores on clean data).

5.3.5 Conclusion

We evaluated the VD task of 11 PDA systems using ten realistic noise conditions controlled in nine SNR levels on two reference databases. We also evaluated these PDAs and the proposed system on a corpus of spontaneous speech recorded in natural settings. The proposed system introduces a quality increment of 2% on clean data (reaching an accuracy of 0.96) and 8% on global accuracy (clean + noisy data) compared to the best-performing PDA.

This section showed how the Voicing Decision task, essential to obtain reliable f_0 , suffers from a diverse range of noises: how SNR degrades performances, depending on the noise type. Starting from the capabilities of existing PDA algorithms, it is proposed a CNN-based model targeting the specific task of voicing decision, not a complete pitch detection algorithm; the model is more robust to noise than the compared PDAs for all types and noise levels evaluated. In the next section, another important aspect for obtaining reliable f_0 curves, the f_0 tracking, is tackled.

5.4 DYNAMIC PROGRAMMING ALGORITHM

For each timeframe of an audio signal, multiple f_0 estimations are

available. Before anything, specifying that the output measurements have timeframes of 5ms each is useful. Each timeframe has one estimated value per PDA algorithm, i.e., 14 f_0 estimated values per timeframe. A dynamic programming algorithm was implemented based on Weenink (2022), Talkin et al. (1995), and Bartošek (2011). The following explanation is also broadly based on and adapted from these authors, but some changes to their original implementation have been made. PDAs often produce very deviant estimations (high standard deviation) and sometimes very coherent estimations (small standard deviation). The deviant estimations often occur in zones with high noise levels, creaky voices, or where no f_0 should be perceived. Coherent estimations occur where the signal is clearer and f_0 can be more easily estimated.

We deal with each PDA estimation for each timeframe as if it is an *f_0 candidate*, since only one estimation will be picked up at the end of this procedure. We must now decide what the best candidate is. For this decision, we will make three assumptions. The first assumption is that each algorithm will make its best bet, i.e., it will send out its most likely f_0 estimation for each timeframe, considering its strongest candidates and other post-treatment procedures. Observing the timeframes where f_0 estimations are highly coherent, we are led to think that the real-world f_0 value must be most likely situated where most PDAs agree. In other words, the closer to the median value (always for the same timeframe), the more likely an estimation is. We use this as a heuristic for all timeframes, even if they display deviant estimations. This is our first assumption: the real-world value is close to where most algorithms agree. At first sight, this may look problematic, but very deviant timeframes will likely have their estimations zeroed when we apply our voicing decision model. We will talk about this later. For now, we will be focused on f_0 tracking. So, to sum up, for each timeframe, the closer to the median a PDA estimated value is, the stronger it will be. The second assumption is that each algorithm has a different sensitivity to different phenomena (resilience to noise, for instance) and that their fortes should be reflected in the final strength of their bets. For now, suffice it to say that we still need to empirically find the reward and penalty factors to be globally applied to the costs of each PDA's estimation. That said, we can now

show the *within-frame cost function* that accounts for the fact that the more distant from the median a PDA estimation is, the more costly its estimation will be. This cost function is based on Bartošek (2011). The equations below give the probability of $a(x)$ for each PDA's f_0 estimation for timeframe t (f_{0k}) with respect to the median value of f_0 estimations of all PDA algorithms (Med_t) in the same timeframe:

Equation 3 – Semitone cents between two frequencies

$$x = 1200 \left| \log_2 \left(\frac{Med_t}{f_{0k}} \right) \right|$$

Equation 4 – Within-frame probability

$$a(x) = \frac{1}{e^{0.0012x}}$$

The cost function within the frame constitutes only a partial solution, as its validity is confined to the consideration of each timeframe in isolation. Notwithstanding certain specific cases, the vocal folds produce vibrations characterized by varying frequencies, whether increasing, remaining constant, or decreasing continuously. Consequently, real-world fundamental frequency (f_0) values exhibit context-dependent behaviors. In cases deemed exceptional, it is assumed that the majority or nearly all of the Pitch Detection Algorithm (PDA) estimations will likely indicate abrupt changes – marking the third and final assumption.

Selecting the strongest f_0 estimation (i.e., the one closest to the median) at a local level (within each timeframe) does not consistently yield the genuine global f_0 contour in the real world – the contour formed when considering the estimated values across all timeframes. Indeed, opting for the strongest local values may result in a highly discontinuous global f_0 contour. At this point, a heuristic to determine

which f_0 estimation to choose when their strengths are equal needs to be addressed.

To address these issues, a function that incorporates the costs associated with transitioning from one f_0 estimation to the next is required. The objective is to ensure a smooth f_0 contour without it being entirely flat. This is achieved by discouraging substantial f_0 changes between timeframes, constituting the between-frames cost. The between-frame cost function is formulated below, resembling the within-frame cost function, with the distinction that the starting point for determining the transition probability $a(x)$ is not the median value in timeframe t , but rather an f_0 estimation in timeframe t and an f_0 estimation in timeframe $t+1$:

Equation 5 – Semitone cents between two frequencies

$$x = 1200 \left| \log_2 \left(\frac{f_{0_t}}{f_{0_{t+1}}} \right) \right|$$

Equation 6 – Between-frames probability

$$a(x) = \frac{1}{e^{0.0012x}}$$

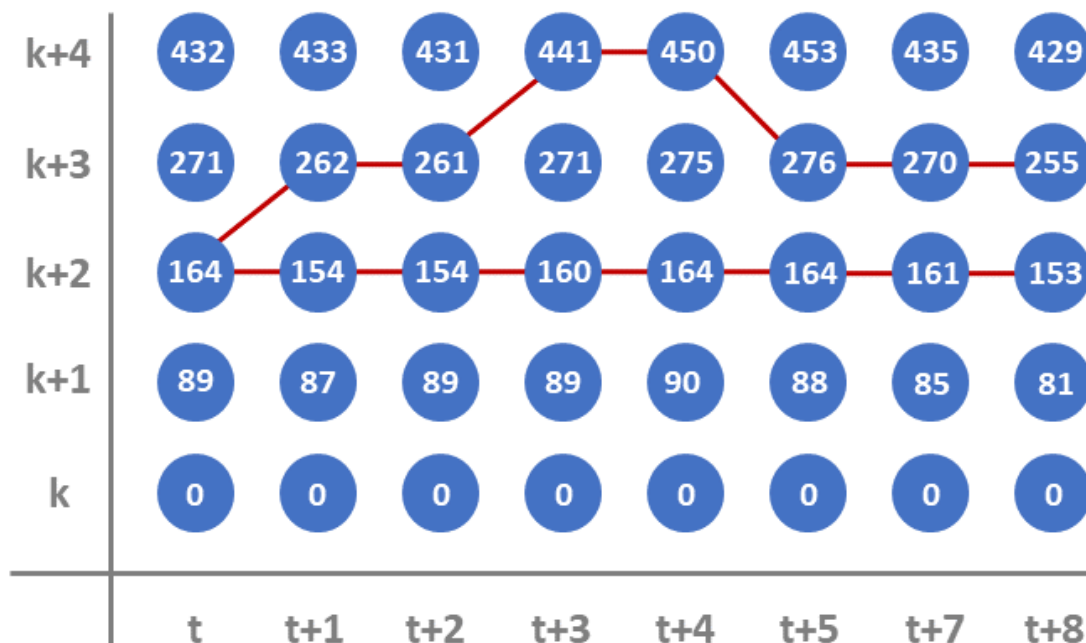
Moreover, transition costs do not come into play for candidates with equal frequencies; costs are only incurred when there is a change in frequency.

There is also a consideration for introducing a penalizing factor to restrict changes from one PDA to another. The objective is to preserve an algorithm's estimations to the greatest extent possible. The intention is not to simply flatten the global fundamental frequency (f_0) contour by consistently selecting the next f_0 point with the least steep inclination from one frame to the next, without considering

which Pitch Detection Algorithm (PDA) produced it. Stability is sought to a considerable extent, aiming to maintain contours. This factor is designed to, to some degree, safeguard the contours generated by a particular algorithm.

With that being said, the optimal f0 tracking is the track that incurs the minimum global cost, accounting for within-frame costs, between-frame costs, and penalties and rewards applied to PDAs. Numerous algorithms are available for finding optimal, least costly tracks – or, succinctly, a *path*. A path in this context refers to a sequence connecting f0 candidates in successive timeframes. The start point may be chosen as the f0 candidate displaying the maximum agreement with the median value in the first frame, and the end can occur at any candidate in the last frame. The figure below illustrates two paths over eight consecutive timeframes, presenting estimations from five PDA algorithms for each timeframe.

Figure 22 - F0 trellis



This figure depicts a trellis, where each path corresponds to a possible global f_0 contour assignment. Having m PDAs' estimations by timeframe and n timeframes, we arrive at a total m^n possible paths – time complexity $O(m^n)$. Even with a relatively small number of PDAs and audio of short duration, the number of possible paths to walk through is very high from a computational standpoint. To manage this task, an algorithm that narrows down the number of steps to be performed by taking on a few assumptions is needed. The Viterbi algorithm (Viterbi, 1967) is a dynamic programming algorithm widely used both for estimating the Maximum a Posteriori Probability estimate of the most likely sequence of hidden states (Hidden Markov Models) and for finding the optimal path through a chain of nodes/events having a cost function. It is also widely used in computational linguistics applications like speech recognition, speech synthesis, diarization, and keyword spotting. It can cut the exponential time complexity from $O(m^n)$ to $O(n*m^2)$. For that 600-ms audio, the number of operations is greatly reduced, and the total time to obtain a result is less than 1 ms.

The underlying assumption is that the most likely path from the first time point up to a given time point t must depend only on the f_0 estimations of timeframe t and the most likely sequence of f_0 points that led to that state at timeframe $t-1$. Put simply, the algorithm only takes into account neighboring frames. The probabilities or costs are evaluated locally, and there is no explicit dependence on timeframes with more than one timeframe behind. It follows that the calculation must be executed sequentially, the path going always in the same sense of time – from left to right, in a spatial representation.

The Viterbi algorithm operates on the state machine assumption. That is, at any time point, the f_0 point being modelled is chosen from a finite number of states. Each state is given by a PDA's f_0 estimation and its within-frame and between-frame costs. Multiple chains of states (paths) lead to a certain state, but only (or at least) one of them is the most likely path to that state because it entails the least costly path, also known as the *winning path* or the *Viterbi path*. The algorithm does not keep track of all possible paths and costs associated with leading to a certain state, as a complete solution would

do. It will examine all local transitions leading to a state (from state $s-1$ to s) and keep the most likely one. The index of the most probable, least costly state is stored, and the next timeframe will be evaluated. This is a key assumption of the Viterbi algorithm. A second key assumption is that a transition from a state to the next one entails transition probabilities or transition costs. The transition costs are computed from within-frame and between frame costs. The last assumption is that a cumulative cost can be achieved by summing the state-to-state transition costs. The algorithm stores the cumulative costs in each state. Then, it goes forward and combines the cumulative costs of all possible previous states with the local transition costs. The algorithm evaluates the combinations of local costs and accumulated costs and pick up the least costly transition. All other paths are discarded. When the end of the trellis is reached, we will have the lowest accumulated cost (the maximum accumulated probability), the chosen f_0 estimation at each frame and the chosen PDA index frame by frame.

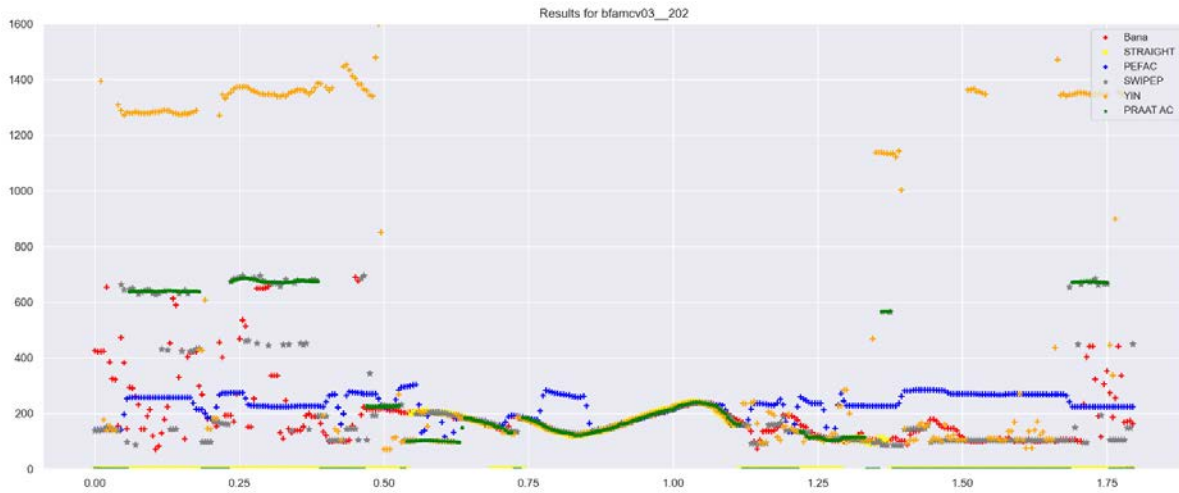
5.5 OUTPUT

5.5.1 Viterbi Algorithm

Figure 23 exhibits the f_0 estimations of six PDA algorithms for the audio file `bfamcv03_202`. For the sake of exemplification, only six algorithms that outputs continuous estimations are used here. The estimations of each PDA are color-coded:

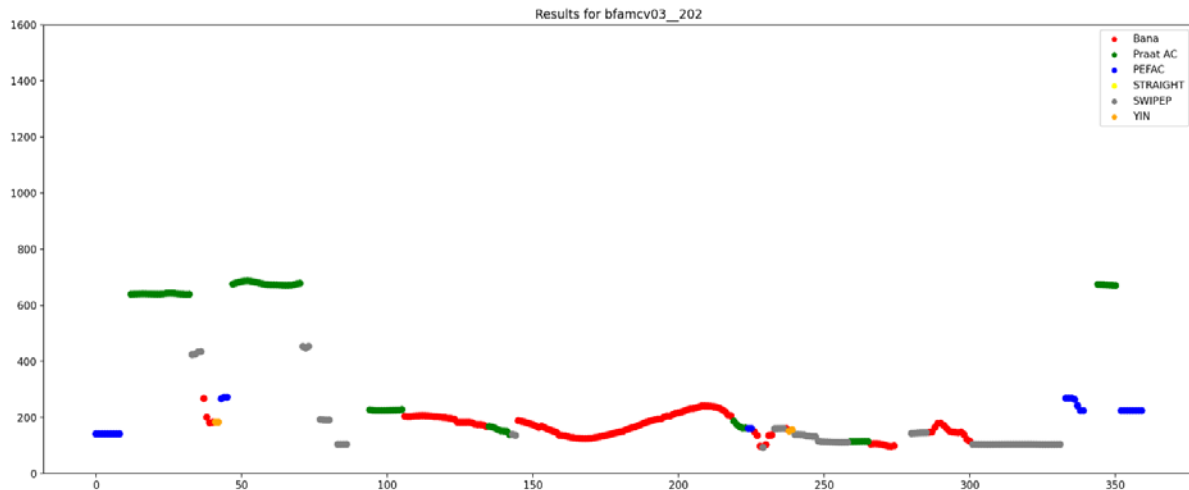
Figure 23 - Raw F_0 estimations of six different PDAs for audio file

bfamcv03_202



It is noteworthy that even in the reliable voiced zone (approximately between 500 and 1250ms), some algorithms produce octave and fifth jumps – see Praat AC and PEFAC. After the data are passed through the VA algorithm, we end up with the winning (or Viterbi) path shown in Figure 24 - Viterbi Path. In this figure, there is only one choice by frame. The chosen algorithm for each frame is color-coded (right upper legend box).

Figure 24 - Viterbi Path for audio file bfamcv03_202



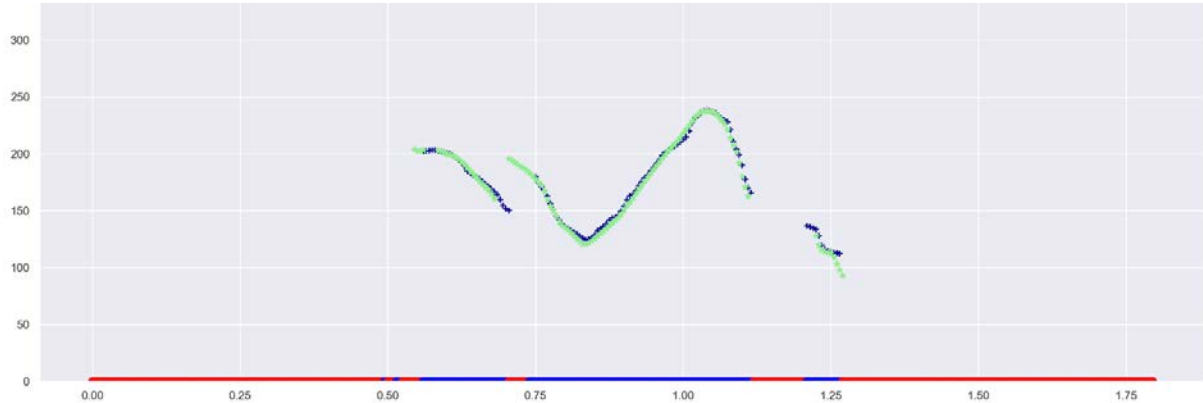
Unduly voiced frames can now be devoiced with the Voicing Decision model.

5.5.2 Voicing Decision Model

Figure 25 displays the predictions of the CNN Voicing Decision model. For visualization purposes, unvoiced time points are zeroed. Unvoiced frames are color-coded in red on the x-axis and voiced frames in blue. F0 estimations are coded in blue and green. Green points represent the f0 estimations passed through the VA algorithm, and blue points represent the ground truth (unseen data) used to evaluate the VD model. Here, the y-axis scale is changed to make the contour clearer.

Figure 25 - Predictions of the best voicing decision model for audio file

bfamcv03_202



As can be observed, there are only two small zones of misassigned voicing decision around 500ms (0.50s), and the model performs well even in voiced zones where PDAs' estimations are more coherent, like in the unvoiced zone between 1000 and 1250ms (check against Figure 23).

5.6 ALGORITHMS USED FOR THE ESTIMATION F0 PARAMETERS

Only six PDA algorithms were effectively used to estimate f_0 through the VA plus VD model solution. They are BaNa (Yang et al., 2014), Praat AC (Boersma, 1993), Pefac (Gonzalez & Brookes, 2014), Straight (Kawahara et al., 2005), Swipep (Camacho, 2007), and YAAPT (Kasi, 2002). They were chosen because they yield the best results for the VD modeling task, thus minimizing VDE, and because they represent a good subset of different types of PDAs, with different robustness. After VA plus VD processing, f_0 features were calculated in accordance with Section 4.4. (Prosodic-acoustic parameters estimation).

In the next chapter, descriptive statistics and an exploratory data analysis are presented.

6 DESCRIPTIVE AND INFERENCE STATISTICS OF THE DISCOURSE MARKERS

This chapter presents the descriptive statistics of the prosodic-acoustic features employed for the classification model (Chapter 7) and an Exploratory Data Analysis (EDA – Tukey, 1977). The EDA is applied to summarize the primary characteristics of the descriptors within a dataset. Utilizing data visualization techniques such as boxplots, histograms, and scatterplots, EDA aims to provide insights into hypotheses, qualitative analyses, and potential errors (outliers). The primary objective is to visually examine what is revealed about the set of descriptors, including means, medians, distributions, skewness, variance, and covariance. The 30 features listed in Chapter 4 are described, each accompanied by the following summary information:

- a) Arithmetic mean;
- b) Standard deviation;
- c) Median;
- d) Trimmed mean;
- e) Minimum value;
- f) Maximum value;
- g) Range;
- h) Asymmetry;
- i) Kurtosis.

The distribution of each feature is illustrated through boxplots for each Discourse Marker function. A Kruskal-Wallis non-parametric test was

applied to compare differences between pairs of Discourse Markers for each feature. A significance level of 0.05 was employed. A table displaying significant differences between pairs of DM categories for each feature is presented after the boxplot. Interactions between features within each group are visualized through pairplots that incorporate the feature's histogram (on the diagonal) and Pearson's correlation coefficients (PCC) between pairs of features.

All statistical summaries, analyses, and plots were generated using Python (Matplotlib, Seaborn, and Scipy Stats).

6.1 FEATURES OF INTENSITY

The set of intensity features encompasses mean intensity (`mean_intensity_dm`), standard deviation of intensity (`std_intensity_dm`), maximum intensity (`max_intensity_dm`), minimum intensity (`min_intensity_dm`), mean intensity on the stressed vowel (`mean_intensity_stressed_dm`), and spectral emphasis on the stressed vowel (`mean_se_stressed_dm`). These features were individually estimated and normalised for each Discourse Marker (DM) instance relative to its node COM.

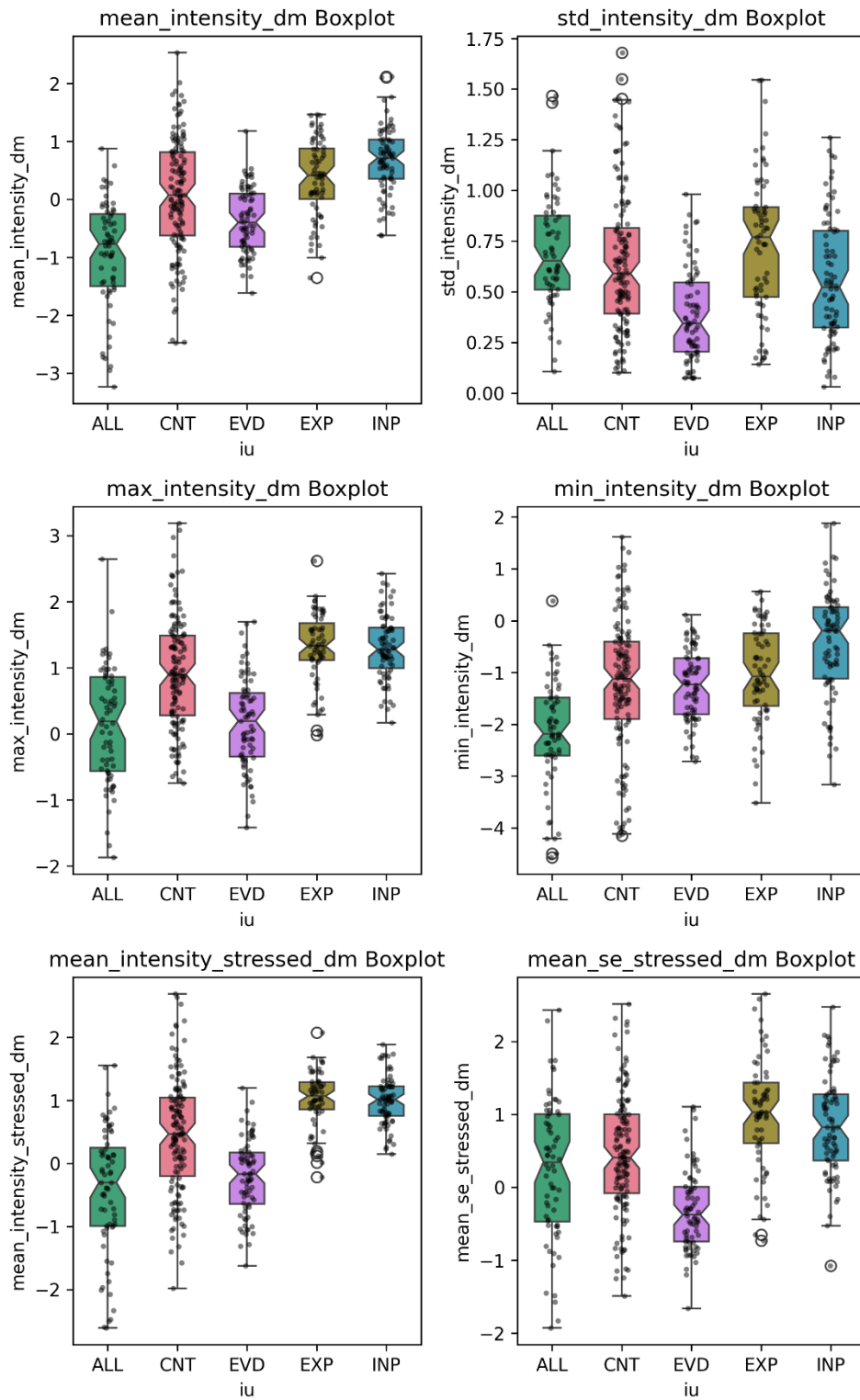
In Figure 26 below, the distribution of each feature is presented by DM function, color-coded to match the hues used in the tables of summary statistics. The boxplots include a notch; non-overlapping notches suggest evidence (at a 95% confidence level) of significantly different medians when comparing boxplots of different DM functions, assuming normal distributions in the compared classes. A Kruskal-Wallis test, summarized shortly after the boxplots, provides a more reliable assessment.

The DM functions tend to follow the order INP > EXP > CNT > EVD > ALL for their mean intensity. This aligns with previous observations by Raso & Vieira (2016) and Gobbo (2019), considering their assessed classes, where the intensity order was INP > CNT > ALL. However, when considering only the stressed vowels (for both mean intensity and spectral emphasis), a different class emerges at the top:

EXP > INP > CNT > EVD > ALL.

Gobbo's model identified intensity features as crucial for distinguishing ALL, CNT, and INP. In his sample, ALL consistently occupied the final position relative to the COM, CNT could be in any position, and INP was consistently in the initial position. There is a noticeable trend for intensity in DM functions in the initial position to be higher than in the final position. This trend aligns with the natural expectation that near terminal boundaries, segments are elongated, and f_0 and intensity decrease. Despite this tendency, a notable contrast can still be observed in the final position between classes with generally higher intensity (CNT and EVD) and ALL, the latter displaying the lowest intensity among all classes.

Figure 26 - Distribution of features of intensity by class of DM



The table presented below provides a summary of the Kruskal-Wallis tests conducted on pairs of Discourse Markers classes. An initial observation based solely on mean intensity might suggest distinct distributions across all DM functions. However, upon closer examination, it becomes apparent that the differences between EXP and INP and ALL and EVD do not reach statistical significance when exclusively considering the stressed vowel. Focusing on measures of the stressed vowel may offer a more reliable indicator of the volume perceived by interlocutors, as it partially mitigates variations introduced by surrounding segments (given that intensity in consonantal segments tends to exhibit greater variability than in vowels). Nonetheless, intrinsic vowel intensity contributes some variation to the system. The absence of a significant difference between EXP and INP appears to be further supported by spectral emphasis.

Table 19 - Significant differences between pairs of DMs by feature of intensity

DM PAIR		Mean Intensity (DM)	SDT Intensity (DM)	Max Intensity (DM)	Min Intensity (DM)	Mean Intensity (stressed syllable)	Mean Spectral Emphasis (stressed syllable)
CNT	EXP	✓	✗	✓	✗	✓	✓
CNT	ALL	✓	✗	✓	✓	✓	✗
CNT	INP	✓	✗	✓	✓	✓	✓
CNT	EVD	✓	✓	✓	✗	✓	✓
EXP	ALL	✓	✗	✓	✓	✓	✓
EXP	INP	✓	✓	✗	✓	✗	✗
EXP	EVD	✓	✓	✓	✗	✓	✓
ALL	INP	✓	✓	✓	✓	✓	✓
ALL	EVD	✓	✓	✗	✓	✗	✓
INP	EVD	✓	✓	✓	✓	✓	✓

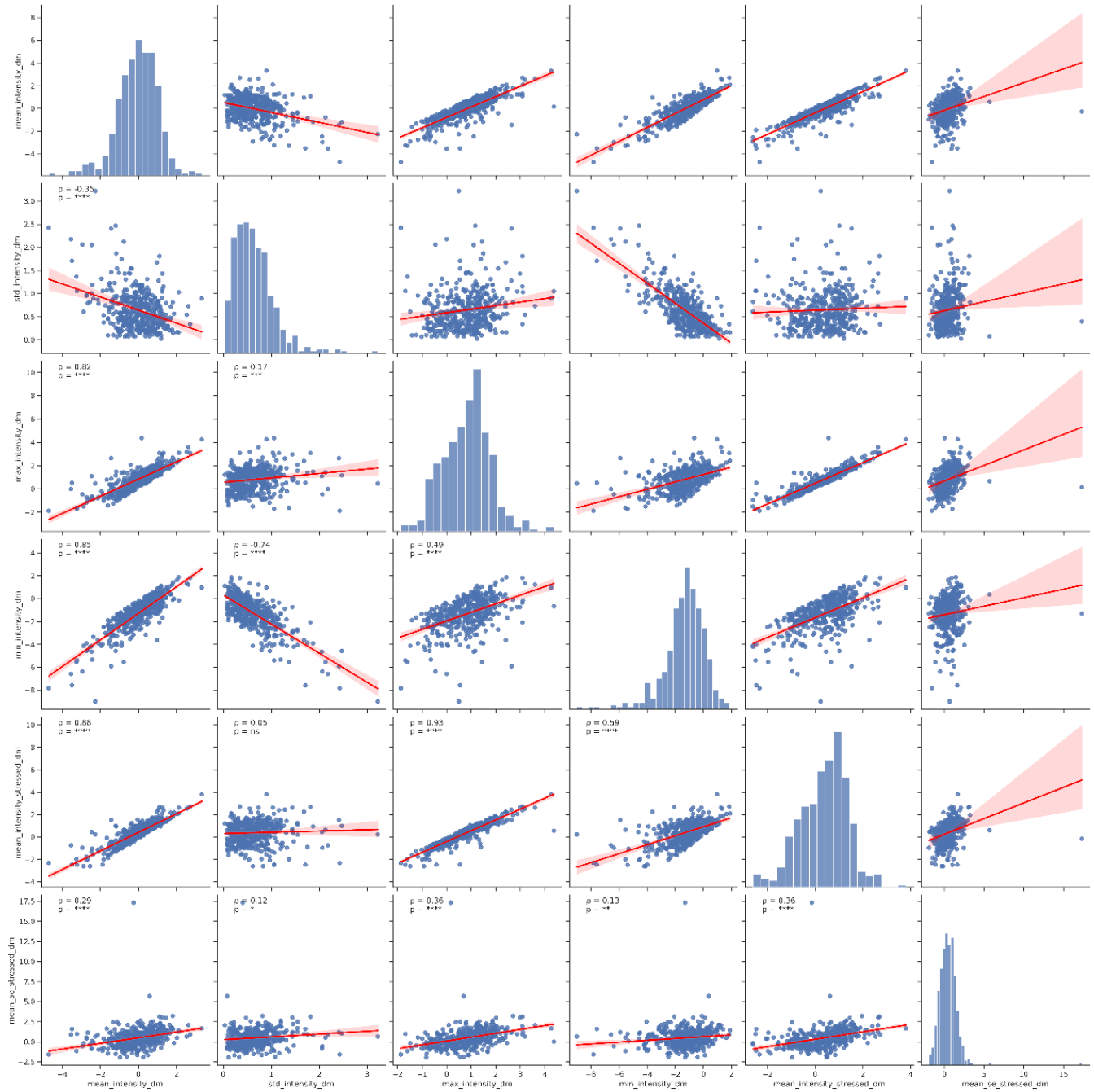
(√ = significant difference; X = non-significant difference)

Table 20 - Statistical summary of the features of intensity

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
Mean Intensity (DM)	ALL	-1.038	1.073	-0.799	-0.950	-4.708	0.876	5.584	-0.984	3.953
	CNT	0.008	1.074	0.048	0.052	-3.542	2.531	6.073	-0.528	3.493
	EVD	-0.367	0.558	-0.397	-0.372	-1.612	1.180	2.792	0.136	2.457
	EXP	0.333	0.718	0.405	0.374	-1.829	1.468	3.297	-0.601	3.039
	INP	0.708	0.796	0.724	0.690	-2.264	3.326	5.590	0.037	5.905
SDT Intensity (DM)	ALL	0.780	0.443	0.684	0.716	0.106	2.425	2.319	1.949	7.511
	CNT	0.706	0.445	0.614	0.652	0.100	2.468	2.368	1.364	5.120
	EVD	0.406	0.251	0.348	0.382	0.073	1.091	1.018	0.868	3.113
	EXP	0.727	0.349	0.769	0.718	0.141	1.545	1.404	0.172	2.474
	INP	0.611	0.457	0.531	0.556	0.031	3.221	3.191	2.855	15.565
Max Intensity (DM)	ALL	0.147	0.878	0.191	0.154	-1.868	2.646	4.514	0.033	2.811
	CNT	0.932	0.920	0.899	0.903	-1.607	4.366	5.972	0.385	3.780
	EVD	0.174	0.689	0.197	0.168	-1.414	1.699	3.113	0.050	2.536
	EXP	1.260	0.721	1.331	1.297	-0.766	3.081	3.847	-0.476	3.823
	INP	1.399	0.709	1.298	1.341	-0.230	4.257	4.487	1.271	6.370
Min Intensity (DM)	ALL	-2.495	1.556	-2.225	-2.305	-7.816	0.380	8.196	-1.471	5.481
	CNT	-1.379	1.503	-1.223	-1.285	-6.558	1.619	8.177	-0.736	3.621
	EVD	-1.243	0.692	-1.229	-1.229	-2.715	0.111	2.826	-0.109	2.229
	EXP	-1.064	0.997	-1.105	-0.995	-4.100	0.563	4.663	-0.632	3.242
	INP	-0.601	1.514	-0.199	-0.444	-8.982	1.878	10.859	-2.606	14.115
Intensity (Stress)	ALL	-0.427	1.018	-0.303	-0.377	-2.610	1.555	4.166	-0.438	2.560
	CNT	0.427	0.955	0.467	0.434	-2.481	2.689	5.170	-0.180	3.036

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
	EVD	-0.211	0.598	-0.167	-0.208	-1.626	1.193	2.819	-0.074	2.454
	EXP	0.878	0.742	1.011	0.941	-1.499	2.433	3.931	-0.963	4.183
	INP	1.075	0.635	1.022	1.030	-0.447	3.801	4.248	1.249	7.079
Spectral emphasis (Stressed vowel)	ALL	0.255	0.965	0.347	0.272	-1.928	2.428	4.356	-0.157	2.609
	CNT	0.634	1.690	0.427	0.494	-1.492	17.330	18.822	7.090	69.790
	EVD	-0.330	0.578	-0.394	-0.352	-1.887	1.102	2.989	0.241	3.292
	EXP	0.946	0.897	1.010	0.971	-1.329	2.995	4.323	-0.326	3.109
	INP	0.845	0.952	0.804	0.827	-1.614	5.702	7.315	1.319	10.423

Figure 27 - Correlation between features of intensity



The pairplot in figure 25 shows that there is a high collinearity between the mean intensity of the whole DM and that of the stressed vowel.

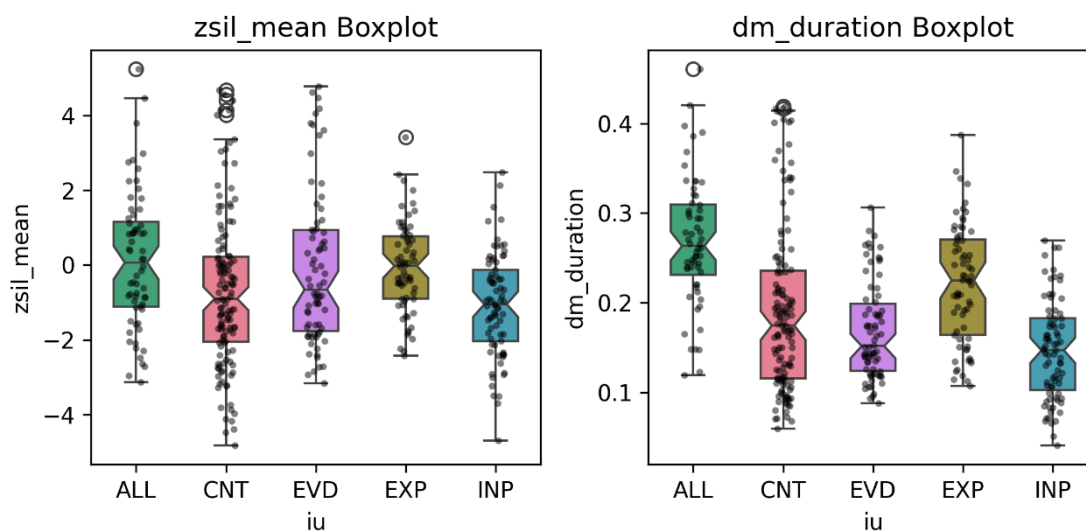
The collinearity is weaker when these two features are compared with the spectral emphasis on the stressed vowel.

6.2 FEATURES OF DURATION

In Figure 28 below, the distribution of duration features is depicted, including the mean standardized duration of syllables (*zsil_mean*) following Barbosa (2013) and the duration of the entire Discourse Marker (DM) relative to COM (*dm_duration*). Notably, two levels of differences are observable: firstly, in the case of ALL and EXP, which exhibit longer durations and tend to have a mean comparable to COM (median around 0); secondly, for CNT, EVD, and INP, which tend to be shorter than ALL and EXP, as well as COM.

Figure 28 - Distribution of the features of duration

Duration Parameters Boxplots



The described tendency is confirmed by the Kruskal-Wallis test, shown below, except for the pair INP-EVD, whose distributions are also significantly different.

Table 21 - Significant differences between pairs of DMs by features of duration

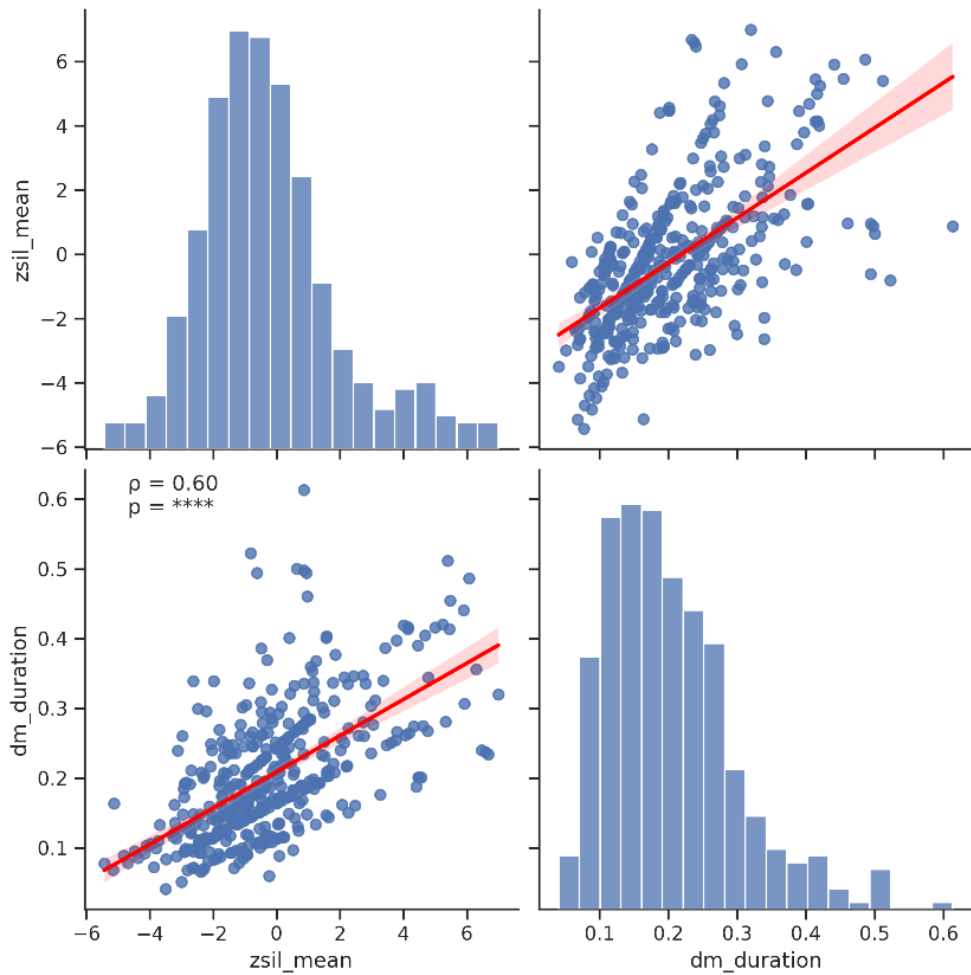
DM PAIR		Mean duration (z-scores)	Total duration (z-scores)
CNT	EXP	✓	✓
CNT	ALL	✓	✓
CNT	INP	✗	✓
CNT	EVD	✗	✗
EXP	ALL	✗	✓
EXP	INP	✓	✓
EXP	EVD	✗	✓
ALL	INP	✓	✓
ALL	EVD	✗	✓
INP	EVD	✓	✓

Table 22 - Statistical summary of the features of duration

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
Articulation rate (DM z-scores)	ALL	0.486	2.297	0.274	0.261	-3.124	6.996	10.120	0.917	3.586
	CNT	-0.335	2.492	-0.725	-0.559	-4.826	6.674	11.500	0.840	3.302
	EVD	-0.017	2.173	-0.552	-0.227	-3.157	5.927	9.084	0.879	2.928
	EXP	0.171	1.608	0.036	0.016	-2.415	6.586	9.001	1.382	6.138
	INP	-1.210	1.576	-1.076	-1.138	-5.436	2.472	7.908	-0.425	3.396
Duration (z-scores)	ALL	0.297	0.105	0.274	0.290	0.119	0.614	0.494	0.891	3.447
	CNT	0.196	0.098	0.176	0.185	0.060	0.455	0.395	0.947	3.047
	EVD	0.171	0.057	0.153	0.166	0.088	0.344	0.256	0.844	2.975
	EXP	0.221	0.066	0.225	0.220	0.107	0.387	0.280	0.122	2.312

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
	INP	0.149	0.059	0.147	0.146	0.041	0.344	0.303	0.591	3.230

Figure 29 - Correlation between features of duration



6.3 FEATURES OF FUNDAMENTAL FREQUENCY (F0)

In Figure 30 below, the distribution of fundamental frequency features is presented. Again, two levels of relevant differences are shown: firstly, for ALL, which exhibits lower fundamental frequency (f0), lower max f0, and lower min f0; and secondly, for CNT, EXP, EVD, and INP, which tend to have a mean f0 approximately at the same level as COM.

An observation regarding INP is relevant here. In previous studies, INP was perceived to have the highest mean f0 level among all DM functions. In the current proposal, instances with functional similarity and a similar f0 form were added to the INP class, specifically those starting the utterance with a flat f0 profile. The height of the tone can vary, with INP displaying a low to medium flat tone or a high flat tone based on the speaker's attitude. The inclusion of instances with lower flat tones in the class is reflected in the mean f0 levels of INP. Nevertheless, it remains the DM function with the highest central tendency of mean f0. However, this difference does not appear significant when comparing INP and EXP for mean f0 and max f0.

Another relevant aspect of the data is the observation that INP exhibits the least spread when considering the standard deviation of f0 points (also, the lowest mean STD in Table 9). This indicates that this DM function has the flattest f0 contour, as expected.

Figure 30 - Distribution of the features of f0

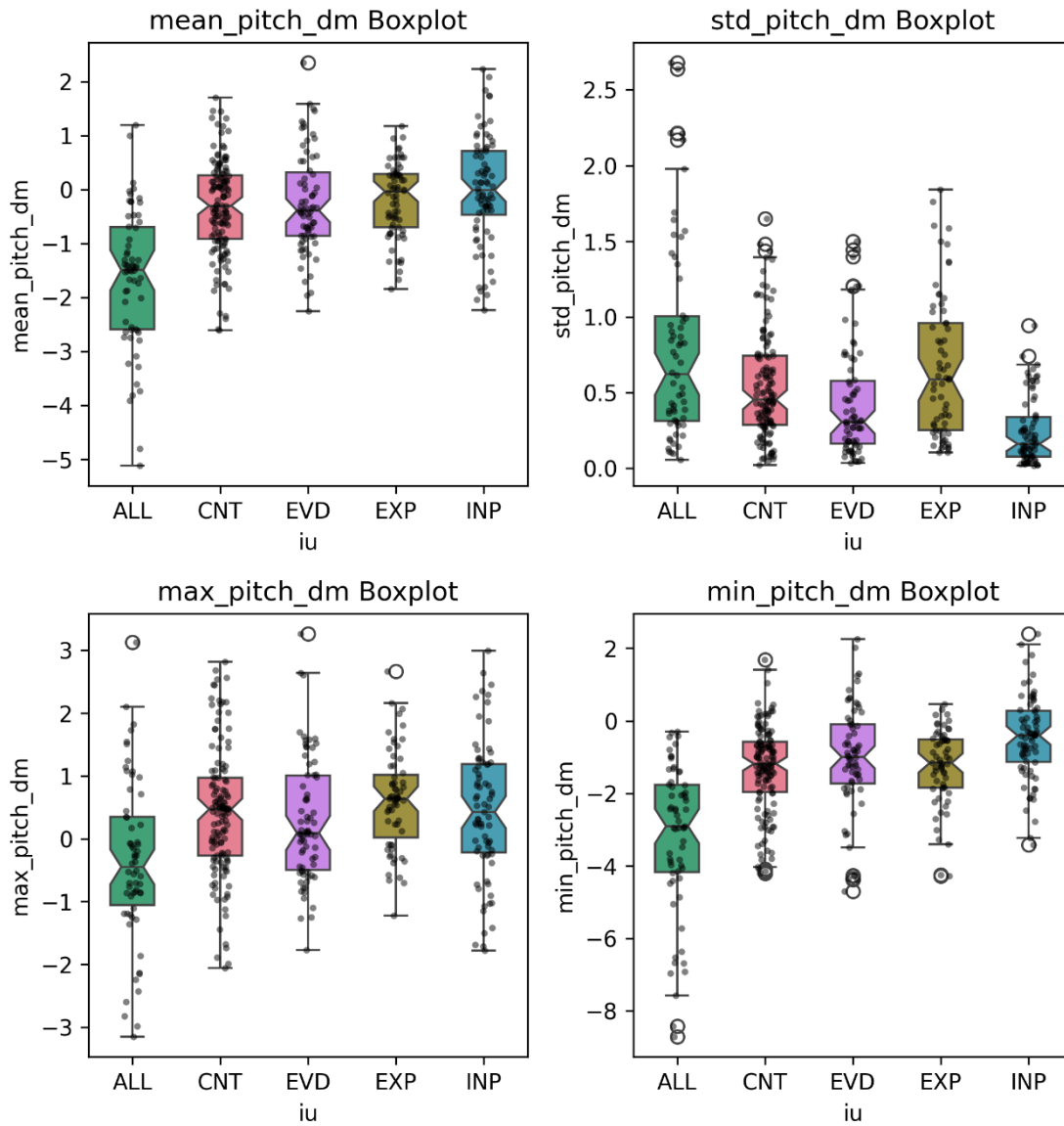


Figure 31 - Significant differences between pairs of DMs by features of f0

DM PAIR		Mean f0	STD f0	Max f0	Min f0
CNT	EXP	X	X	X	X

DM PAIR		Mean f0	STD f0	Max f0	Min f0
CNT	ALL	✓	✓	✓	✓
CNT	INP	✓	✓	✗	✓
CNT	EVD	✗	✓	✗	✗
EXP	ALL	✓	✗	✓	✓
EXP	INP	✗	✓	✗	✓
EXP	EVD	✗	✓	✓	✗
ALL	INP	✓	✓	✓	✓
ALL	EVD	✓	✓	✓	✓
INP	EVD	✓	✓	✗	✓

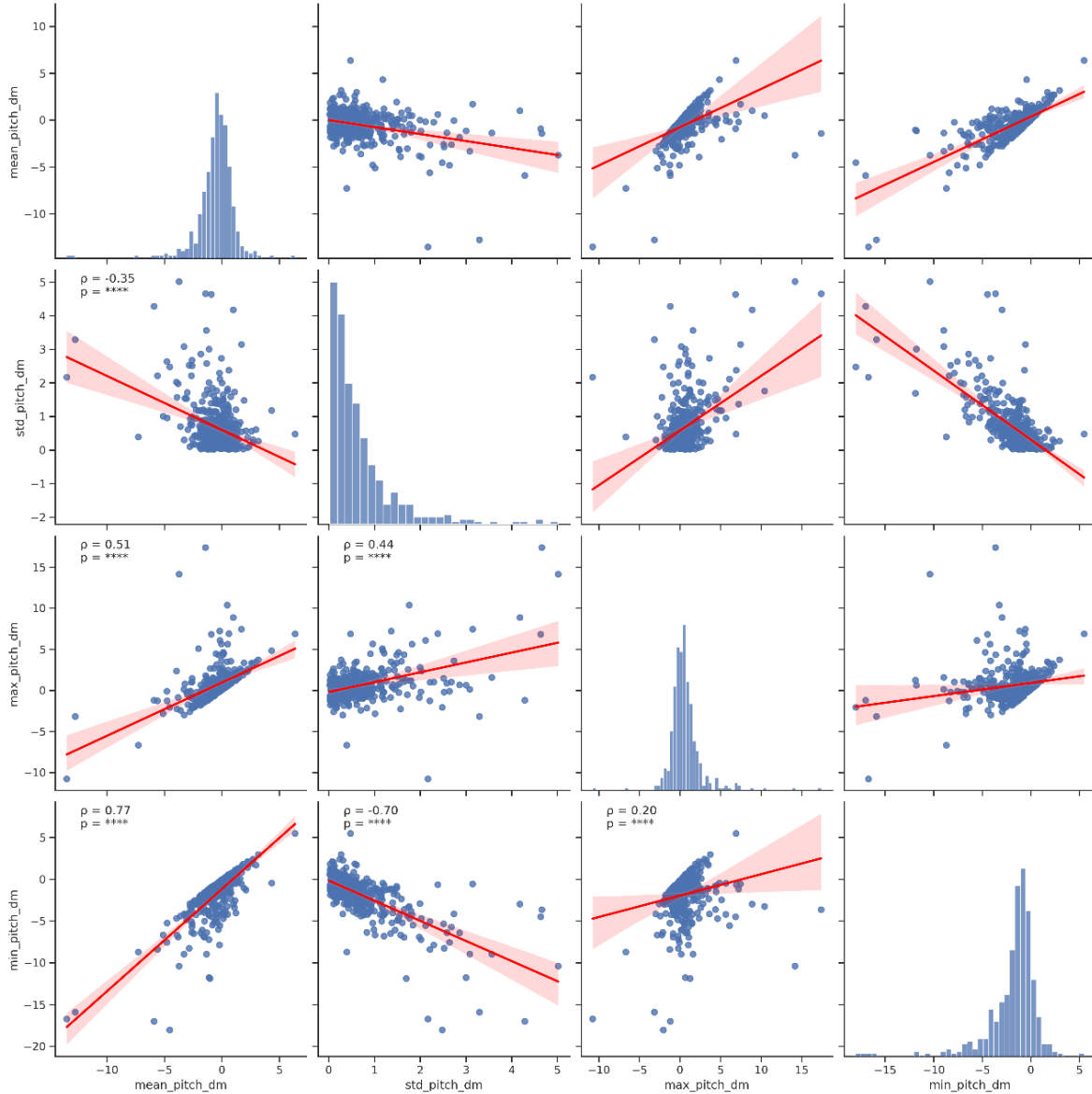
Table 23 - Statistical summary of the features of f0

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
Mean f0	ALL	-2.185	2.486	-1.513	-1.796	-13.516	1.196	14.712	-2.747	12.190
	CNT	-0.351	1.102	-0.307	-0.320	-4.548	2.973	7.521	-0.388	4.663
	EVD	-0.266	1.218	-0.391	-0.274	-4.780	3.180	7.960	-0.154	5.262
	EXP	-0.155	0.732	-0.019	-0.144	-1.845	1.937	3.782	-0.064	2.974
	INP	0.114	1.409	0.041	0.054	-3.950	6.379	10.329	1.084	7.778
STD f0	ALL	1.128	1.201	0.755	0.892	0.058	5.017	4.959	1.855	5.642
	CNT	0.689	0.618	0.496	0.578	0.022	3.145	3.122	1.987	7.036
	EVD	0.512	0.496	0.312	0.423	0.038	2.295	2.257	1.687	5.332
	EXP	0.800	0.719	0.599	0.682	0.107	3.566	3.459	1.839	6.529
	INP	0.343	0.415	0.192	0.255	0.019	2.117	2.098	2.434	9.327
Max f0	ALL	0.102	3.661	-0.375	-0.296	-10.754	17.375	28.129	2.252	12.873
	CNT	0.733	1.555	0.487	0.562	-2.054	7.442	9.497	1.566	6.811
	EVD	0.324	1.053	0.101	0.248	-1.770	3.735	5.505	0.882	3.857
	EXP	1.048	1.775	0.663	0.751	-1.227	10.396	11.623	3.092	14.645
	INP	0.762	1.635	0.542	0.573	-1.777	6.878	8.655	1.502	6.058

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
Min f0	ALL	-4.045	3.627	-2.947	-3.421	-16.998	-0.298	16.699	-2.055	7.199
	CNT	-1.600	2.119	-1.232	-1.349	-18.027	1.835	19.862	-3.992	28.677
	EVD	-1.444	2.146	-1.061	-1.246	-7.888	2.951	10.839	-0.980	4.032
	EXP	-1.834	2.152	-1.253	-1.462	-11.775	0.467	12.241	-2.409	9.678
	INP	-0.607	1.615	-0.435	-0.543	-5.556	5.474	11.030	-0.108	5.774

Figure 32 below shows that there are significant correlations between mean f0, on the one hand, and maximum and minimum f0 levels on the other.

Figure 32 - Correlation between features of f0



6.4 FEATURES OF F0 VARIATION

The features of f0 variation were specifically crafted to capture the general movements of fundamental frequency (f0) with respect to the stressed vowel (Gobbo, 2019). The features include the regression line on the entire Discourse Marker instances (pitch_slope_dm), on the

stressed vowel (*pitch_slope_stressed*), the *f0* range (max – min: *pitch_range_dm*), *f0* slope before the mid-point of the stressed vowel (*pitch_slope_before_stressed_dm*), and *f0* slope after the mid-point of the stressed vowel (*pitch_slope_after_stressed_dm*). While another approach involves fitting a polynomial curve to the data and utilizing the polynomial's coefficients as descriptors, it may not effectively capture the flat *f0* profiles typically found in instances of the INP class. Therefore, we retained the features proposed by Gobbo (2019) for the sake of comparability and as a robust descriptor of the tendencies observed along the stressed vowels.

A strong correlation is observed between the movements on the entire DM and those on its stressed vowel. Consequently, our focus will primarily be on the stressed vowel. The qualitative analysis proposed that the ALL and CNT DMs would exhibit negative slopes along the stressed vowels (falling *f0* movements), EXP and EVD positive slopes (rising *f0* movements), and INP the flattest movements (*f0* slope ≈ 0). These expectations can be confirmed in the boxplots. In terms of the absolute values of *f0* slope, the ascending order from the flattest to the steepest is: INP < EVD < EXP < ALL < CNT (this order is confirmed in the summary statistics table).

Another noteworthy tendency is observed in the EVD class. Despite being perceptually characterized by a distinctive rising movement, the class exhibits the second flattest *f0* movement over the stressed vowel. Most EVD instances occur in the final position of terminated sequences; a region correlated with falling *f0* profiles conveying terminal boundaries. Here, we hypothesize that a sustained, almost flat movement is sufficient to mark the EVD function. The contrast is not with a flat movement but rather with a baseline falling movement.

Figure 33 - Distribution of features of f0 variation

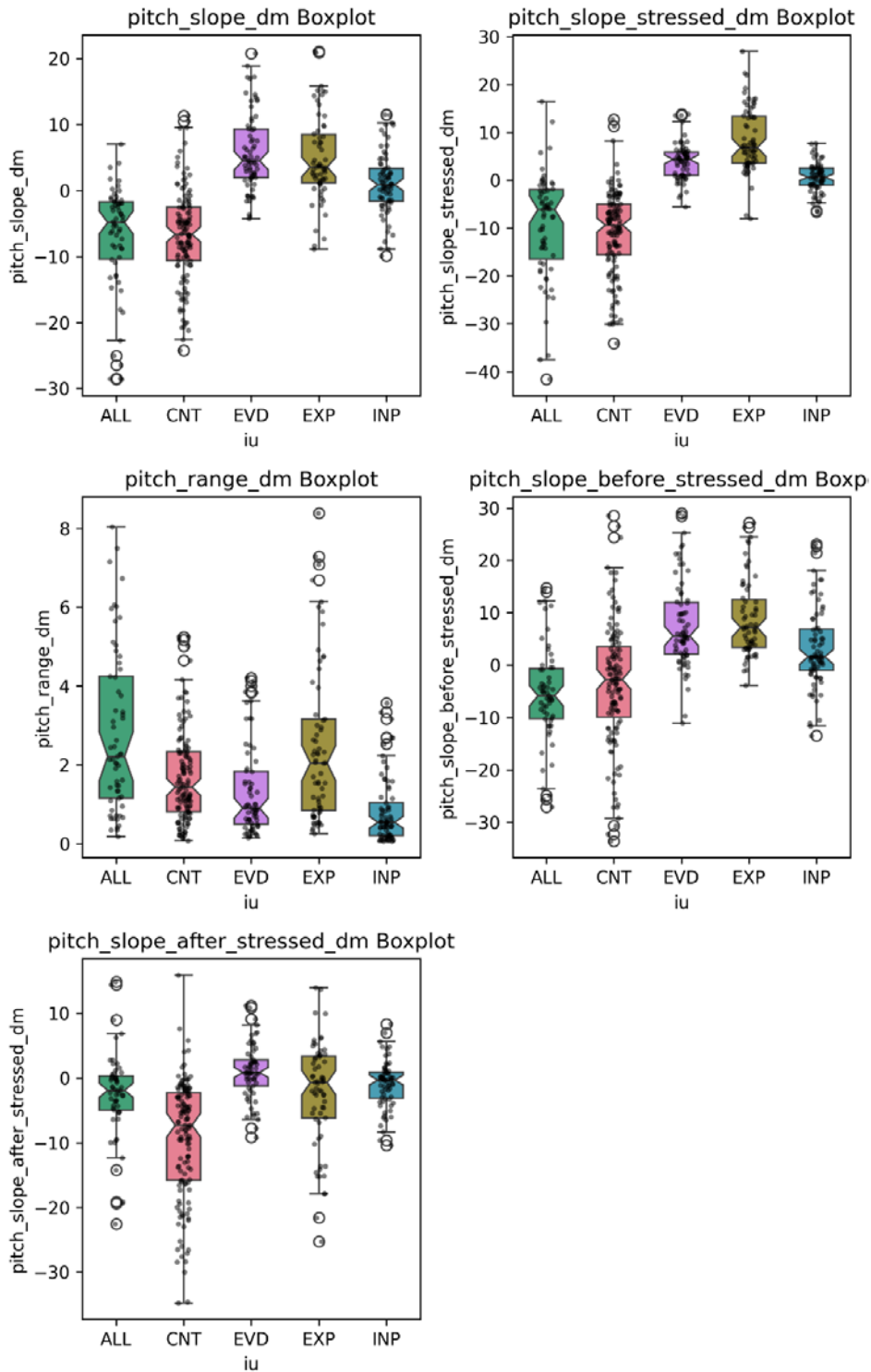


Table 24 - Significant differences between pairs of DMs by features of *f0* variation

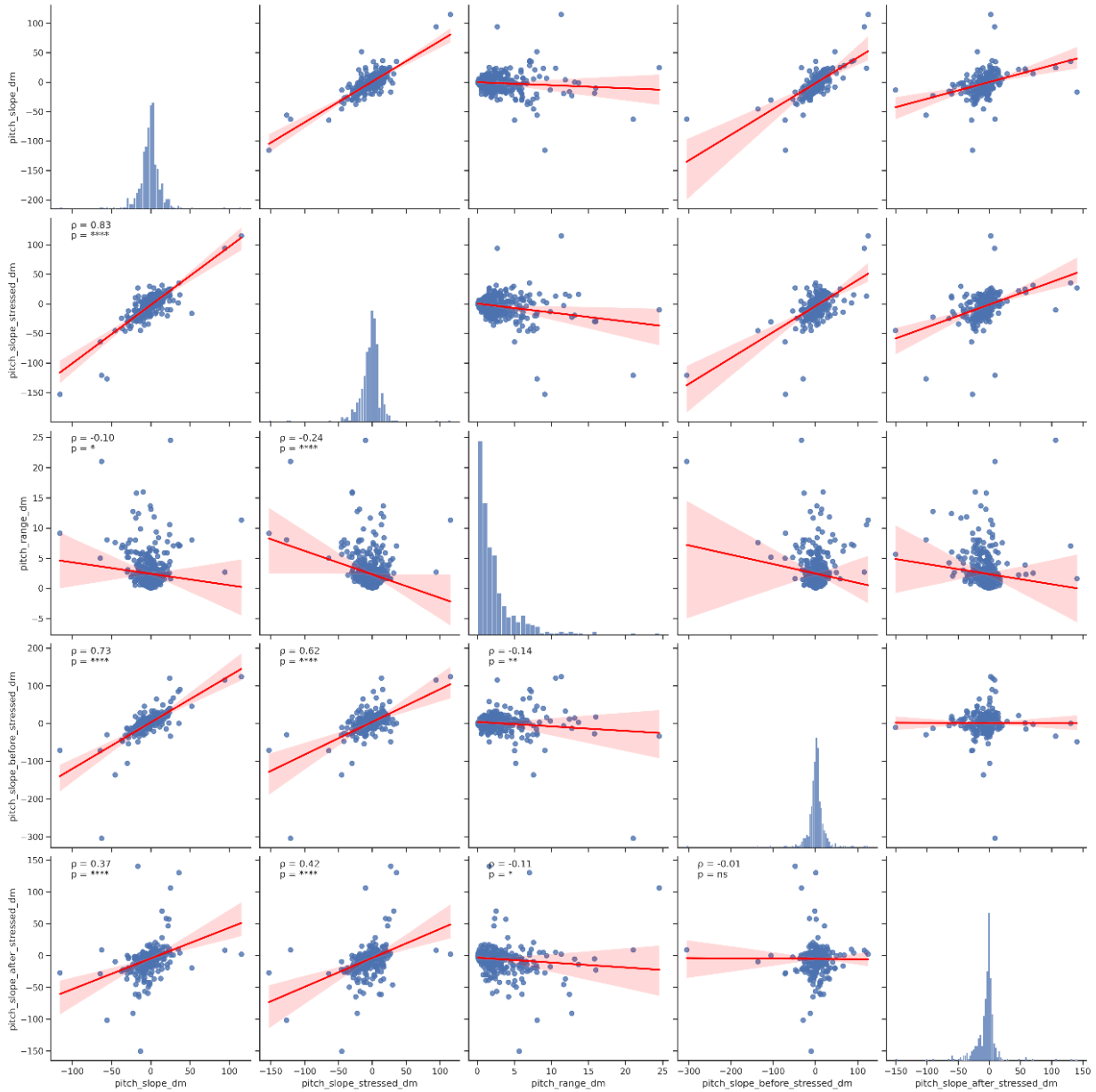
DM PAIR		F0 slope	F0 slope (stressed vowel)	F0 range	F0 slope before stressed vowel	F0 slope after stressed vowel
CNT	EXP	✓	✓	✗	✓	✓
CNT	ALL	✗	✗	✓	✓	✓
CNT	INP	✓	✓	✓	✓	✓
CNT	EVD	✓	✓	✓	✓	✓
EXP	ALL	✓	✓	✗	✓	✗
EXP	INP	✓	✓	✓	✓	✗
EXP	EVD	✗	✓	✓	✗	✓
ALL	INP	✓	✓	✓	✓	✓
ALL	EVD	✓	✓	✓	✓	✓
INP	EVD	✓	✓	✓	✓	✓

Table 25 - Summary statistics of the features of *f0* variation

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
F0 slope (DM)	ALL	-6.927	19.525	-5.190	-7.215	-62.620	115.202	177.822	3.250	24.611
	CNT	-8.122	15.756	-6.824	-7.173	-115.554	52.008	167.562	-2.432	19.880
	EVD	6.428	7.499	4.543	5.860	-14.752	28.682	43.433	0.641	3.856
	EXP	5.355	9.853	3.603	4.920	-19.906	36.860	56.767	0.645	4.771
	INP	3.007	12.782	1.013	1.470	-19.112	94.178	113.289	4.909	34.120
F0 slope (Stressed vowel)	ALL	-9.680	24.336	-6.554	-9.301	-120.815	115.202	236.017	0.541	17.465
	CNT	-13.158	19.359	-9.767	-10.992	-152.867	27.052	179.920	-4.309	29.264
	EVD	4.703	6.759	4.547	4.454	-21.536	24.666	46.201	-0.074	6.522

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
	EXP	8.047	7.141	6.676	8.007	-13.172	26.962	40.134	0.008	3.648
	INP	2.589	13.047	0.645	0.657	-15.662	94.178	109.839	4.853	32.271
F0 range	ALL	4.146	4.682	2.458	3.223	0.182	24.540	24.357	2.439	9.276
	CNT	2.333	2.329	1.607	1.914	0.071	15.973	15.902	2.503	11.581
	EVD	1.767	1.883	0.952	1.409	0.143	8.907	8.764	1.815	5.769
	EXP	2.881	2.818	2.062	2.404	0.258	13.659	13.401	1.905	6.640
	INP	1.369	2.082	0.556	0.879	0.063	11.649	11.586	3.003	12.411
F0 slope before stressed vowel	ALL	-10.514	42.157	-6.445	-7.258	-304.342	124.572	428.913	-4.707	37.409
	CNT	-4.295	22.816	-2.596	-3.128	-135.790	65.403	201.193	-1.972	13.451
	EVD	8.386	12.998	5.649	7.551	-34.210	68.198	102.408	1.172	9.051
	EXP	13.350	21.151	7.253	9.356	-20.559	120.635	141.194	3.293	14.786
	INP	5.459	16.670	1.631	3.400	-30.133	115.431	145.565	4.042	26.066
F0 slope after stressed vowel	ALL	-9.084	28.264	-2.654	-6.170	-150.455	106.521	256.976	-1.427	15.139
	CNT	-9.525	19.513	-7.529	-9.211	-101.661	140.791	242.452	2.624	31.006
	EVD	0.819	12.107	0.890	1.257	-42.434	57.426	99.860	-0.122	12.312
	EXP	-4.961	15.189	-1.602	-2.901	-60.742	28.384	89.126	-2.051	8.524
	INP	0.667	19.745	-0.329	-1.035	-45.402	130.677	176.079	4.139	27.036

Figure 34 - Correlation between features of f0 variation



6.5 FEATURES OF ALIGNMENT

The alignment features were designed to capture the alignment tendencies of the maximum and minimum points of f0 and intensity

within the DM instance, with respect to the central point of the stressed vowel. The boxplots in the first and third rows of Figure 35 shows the position within the DM instance where the maximum and minimum points of intensity and f_0 are achieved. Values closer to 0 indicates that the time point of interest is closer to the beginning of the DM instance whereas values closer to 1 indicates that the time point of interest is closer to the end. The boxplots in the second and fourth rows show how the timepoints of interest are displaced with respect to the central point of the stressed vowels. Values closer to 0 indicate that the timepoint of interest is aligned with the center of the stressed vowel. Values different than 0 indicate a displacement with respect to the stressed vowel central point. The higher the absolute value is, the larger the displacement. Negative values indicate that the timepoint of interest occurs before the central point of the stressed vowel and positive values that it occurs after the central point of the stressed vowel.

Starting with intensity alignment, the maximum intensity tends to be reached shortly after the midpoint of the DM across all classes except for ALL, where the maximum intensity occurs at the DM's beginning. Note that this intensity alignment does not necessarily correlate with other features. For example, the EVD class tends to attain its highest intensity point well before the central point of the stressed vowel. Nevertheless, the peak of the f_0 tends to occur at the end of the unit (compare EVD distributions and medians in the left boxplots of the second and third rows). Conversely, concerning minimum intensity, the tendency is reversed: for all DM functions except ALL, the minimum intensity tends to occur at the beginning of the DM.

In terms of f_0 , the observed distributions in the boxplots correspond to the described curves for each DM function. ALL and CNT display the maximum point at the beginning and the minimum point tending towards the end of the units, indicating a falling f_0 movement. In contrast, EVD and EXP have the minimum f_0 point at the beginning and the maximum point tending towards the end of the unit. INP, however, does not exhibit a clear tendency in this regard. Despite this, the maximum f_0 aligns distinctly with the central point of the stressed vowel.

Figure 35 - Distribution of the features of alignment

Alignment Parameters Boxplots

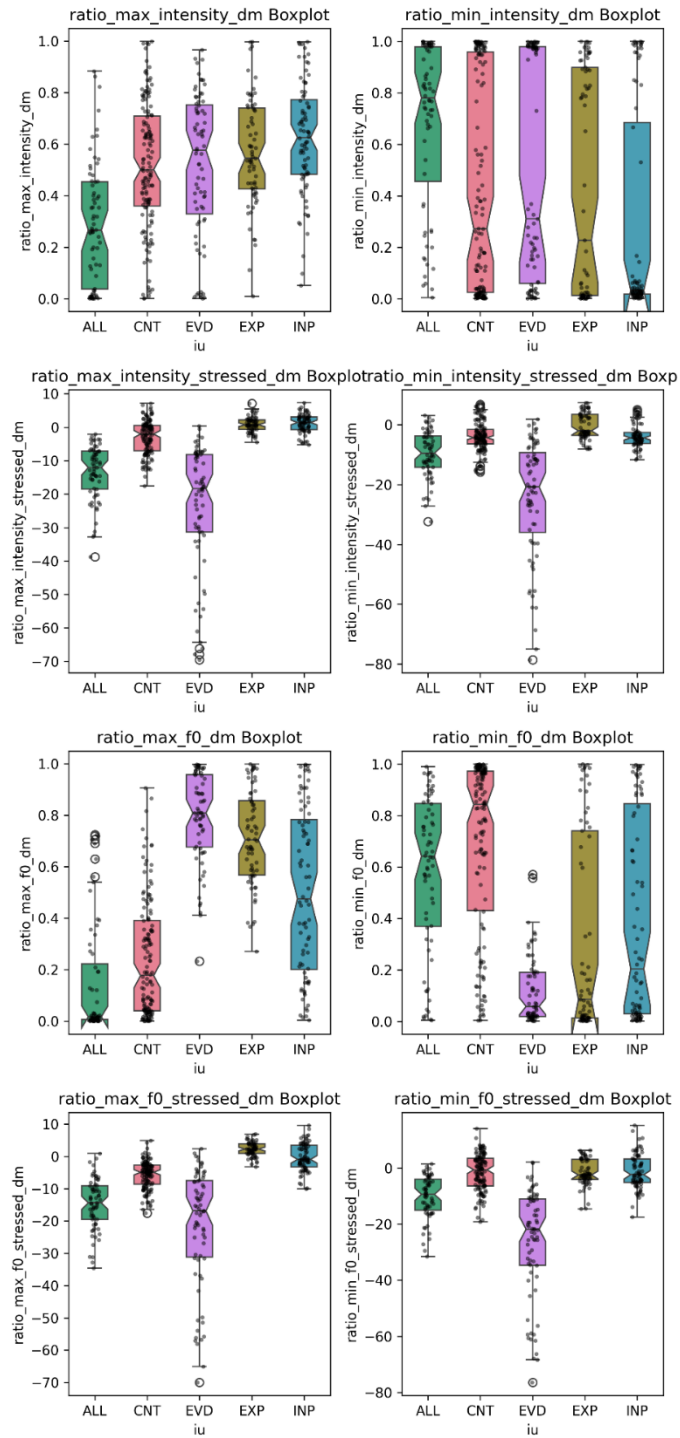


Table 26 shows the significant differences between pairs of DM functions for the alignment features. The ratios of maximum and minimum f0 (inside the DM), and maximum f0 with respect to the central point of the stressed vowel are distinctive across DM functions. This tendency is not so clear when we analyze the coefficients of f0 curves alone, as is shown in the next subsection. Therefore, although these features are reflected in the f0 curves, keeping and testing them in the final classification tasks seems good.

Table 26 - Significant differences between pairs of DMs by features of alignment

DM PAIR		Ratio max Int (inside DM)	Ratio min Int (inside DM)	Ratio max Int (wrt stressed vowel)	Ratio min Int (wrt stressed vowel)	Ratio max f0 (inside DM)	Ratio min f0 (inside DM)	Ratio max f0 (wrt stressed vowel)	Ratio min f0 (wrt stressed vowel)
CNT	EXP	X	X	✓	✓	✓	✓	✓	X
CNT	ALL	✓	✓	✓	✓	✓	✓	✓	✓
CNT	INP	✓	✓	✓	X	✓	✓	✓	X
CNT	EVD	X	X	✓	✓	✓	✓	✓	✓
EXP	ALL	✓	✓	✓	✓	✓	✓	✓	✓
EXP	INP	X	X	X	✓	✓	X	✓	X
EXP	EVD	X	✓	✓	✓	X	X	✓	✓
ALL	INP	✓	✓	✓	✓	✓	✓	✓	✓
ALL	EVD	✓	X	✓	✓	✓	✓	X	✓
INP	EVD	X	✓	✓	✓	✓	✓	✓	✓

Table 27 - Summary statistics of the features of alignment

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
Ratio max intensity (inside DM)	ALL	0.290	0.237	0.266	0.270	0.001	0.882	0.882	0.548	2.600
	CNT	0.515	0.248	0.501	0.520	0.001	0.999	0.998	-0.134	2.324
	EVD	0.521	0.279	0.576	0.536	0.001	0.966	0.966	-0.453	2.095
	EXP	0.567	0.212	0.545	0.568	0.010	0.997	0.987	-0.054	2.790
	INP	0.622	0.217	0.626	0.628	0.051	0.996	0.945	-0.276	2.731
Ratio min intensity (inside DM)	ALL	0.683	0.323	0.780	0.716	0.006	0.999	0.993	-0.856	2.258
	CNT	0.439	0.419	0.272	0.425	0.001	1.000	0.999	0.317	1.320
	EVD	0.509	0.435	0.310	0.511	0.002	0.999	0.997	0.096	1.138
	EXP	0.445	0.434	0.228	0.434	0.001	1.000	0.999	0.139	1.131
	INP	0.275	0.403	0.033	0.218	0.001	0.999	0.998	1.085	2.220
Ratio max intensity (wrt stressed vowel)	ALL	-16.033	15.593	-12.619	-13.585	-114.826	-2.067	112.759	-4.123	25.020
	CNT	-5.098	11.190	-2.339	-3.344	-90.631	7.179	97.810	-4.240	28.489
	EVD	-29.681	33.757	-18.893	-23.542	-203.281	0.364	203.645	-2.935	13.224
	EXP	-0.848	7.891	0.636	0.537	-45.952	7.955	53.907	-4.092	21.017
	INP	-0.352	12.712	1.415	1.193	-103.241	13.228	116.469	-6.957	55.267
Ratio max intensity (wrt stressed vowel)	ALL	-12.471	15.314	-10.180	-10.242	-107.155	3.129	110.284	-3.857	22.907
	CNT	-5.915	11.261	-4.573	-4.458	-86.301	14.029	100.331	-3.741	23.590
	EVD	-29.780	34.213	-20.830	-23.908	-208.517	1.832	210.350	-3.007	14.080
	EXP	-2.022	8.729	-2.244	-0.681	-41.100	7.400	48.500	-2.881	12.909
	INP	-4.748	12.016	-4.214	-3.711	-95.863	11.517	107.380	-5.604	42.707
Ratio max f0 (inside DM)	ALL	0.192	0.288	0.020	0.139	0.000	0.989	0.989	1.523	4.029
	CNT	0.255	0.245	0.186	0.223	0.000	0.968	0.968	1.012	3.319
	EVD	0.724	0.273	0.797	0.769	0.004	0.997	0.993	-1.344	3.957
	EXP	0.700	0.202	0.701	0.712	0.010	0.999	0.989	-0.592	3.526
	INP	0.504	0.323	0.476	0.503	0.005	0.997	0.993	0.030	1.581
Ratio min	ALL	0.582	0.305	0.641	0.601	0.005	0.989	0.984	-0.598	2.131

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
	CNT	0.692	0.334	0.842	0.731	0.004	1.000	0.996	-0.874	2.212
	EVD	0.189	0.264	0.072	0.128	0.000	0.994	0.994	2.113	6.503
	EXP	0.317	0.385	0.085	0.279	0.001	0.999	0.998	0.809	1.854
	INP	0.388	0.391	0.204	0.362	0.001	0.996	0.996	0.468	1.494
Ratio max f0 (wrt stressed vowel)	ALL	-17.141	15.406	-14.585	-15.157	-117.383	0.882	118.265	-4.331	27.642
	CNT	-7.846	10.592	-5.170	-6.134	-92.796	4.945	97.742	-4.749	33.736
	EVD	-27.831	33.454	-17.276	-21.951	-203.063	2.410	205.473	-2.960	13.652
	EXP	0.269	7.745	2.104	1.821	-40.962	6.882	47.844	-3.882	18.568
	INP	-1.709	11.793	-0.957	-0.365	-94.018	14.165	108.184	-6.253	48.405
Ratio max f0 (wrt stressed vowel)	ALL	-13.255	16.078	-10.008	-10.752	-116.653	1.424	118.077	-4.287	26.511
	CNT	-3.519	11.876	-1.360	-1.805	-86.301	14.029	100.331	-3.450	21.004
	EVD	-33.084	34.937	-22.704	-27.169	-209.608	2.042	211.650	-2.825	12.654
	EXP	-3.261	9.052	-2.562	-1.795	-47.893	6.294	54.187	-3.105	14.505
	INP	-3.049	13.232	-2.682	-1.812	-102.626	15.224	117.850	-5.491	41.399

function was selected because it was a good compromise between data fitting and accuracy scores. Furthermore, higher degree polynomials diminish the Mean Square Error (MSE - metric used to evaluate the goodness of fitting) but create unnecessary details, which are probably not perceived by the interlocutor. This choice is further motivated in the next chapter.

Equation 7 - Polynomial coefficients of the cubic function

$$f_0 = coef_0 \cdot x^3 + coef_1 \cdot x^2 + coef_2 \cdot x + coef_3$$

The cubic polynomial has four coefficients. They are encoded, as in Equation 7 - Polynomial coefficients, as coef_0 (1st coefficient), coef_1 (2nd coefficient), coef_2 (3rd coefficient), and coef_3 (4th coefficient). Figure 37 shows the distribution of the coefficients of f0 curve. Emphasis will be placed on the significant differences between pairs of DMs by coefficient.

Figure 37 - Distribution of features of f0 curve

F0 Curve Parameters Boxplots

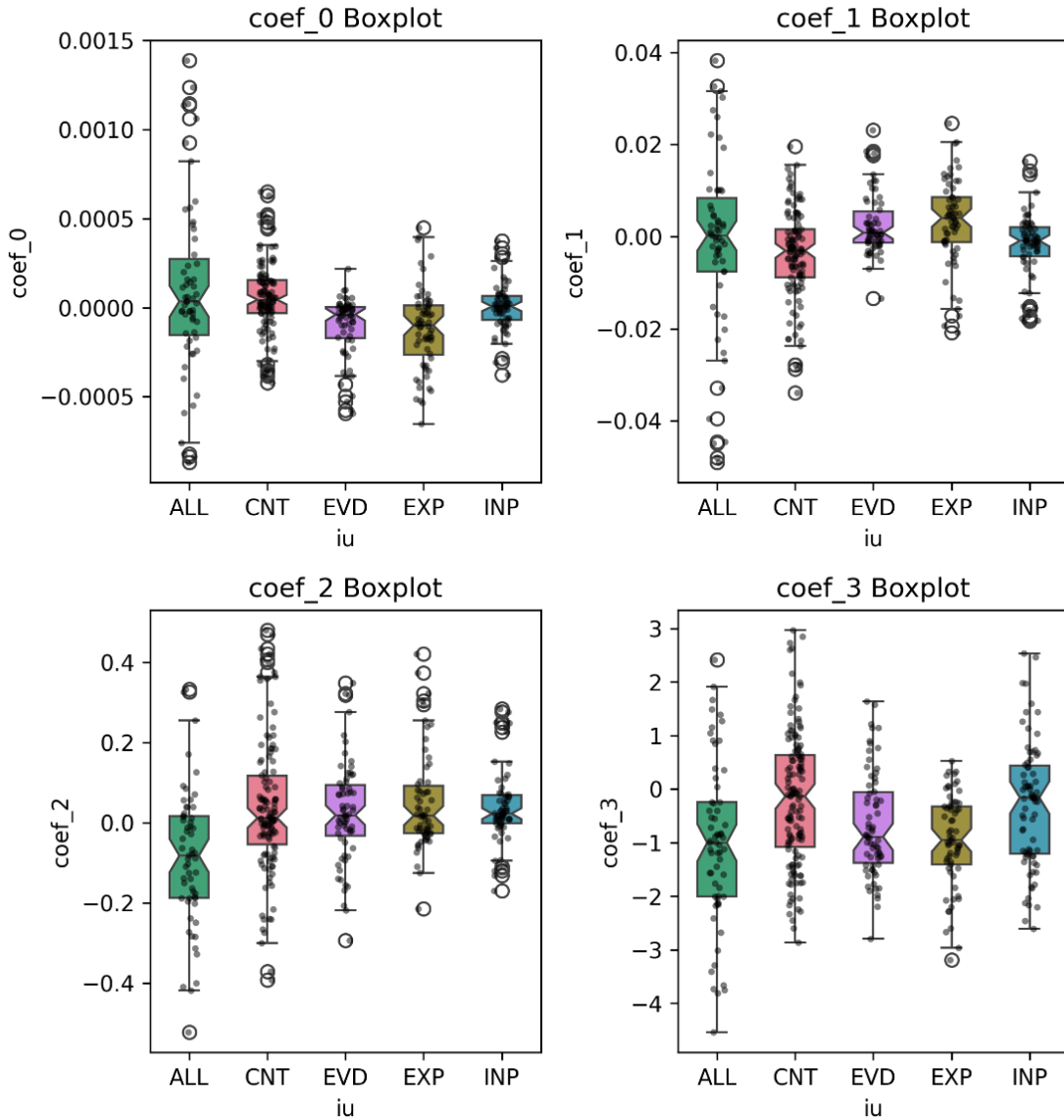


Table 28 shows the pairs of DMs and whether the difference between the distribution of coefficients is significant. For all groups of features analyzed so far, at least one feature always exhibited a significant difference between a pair. Here, there is a special situation. The pairs CNT-INP and EXP-EVD do not display any significantly different

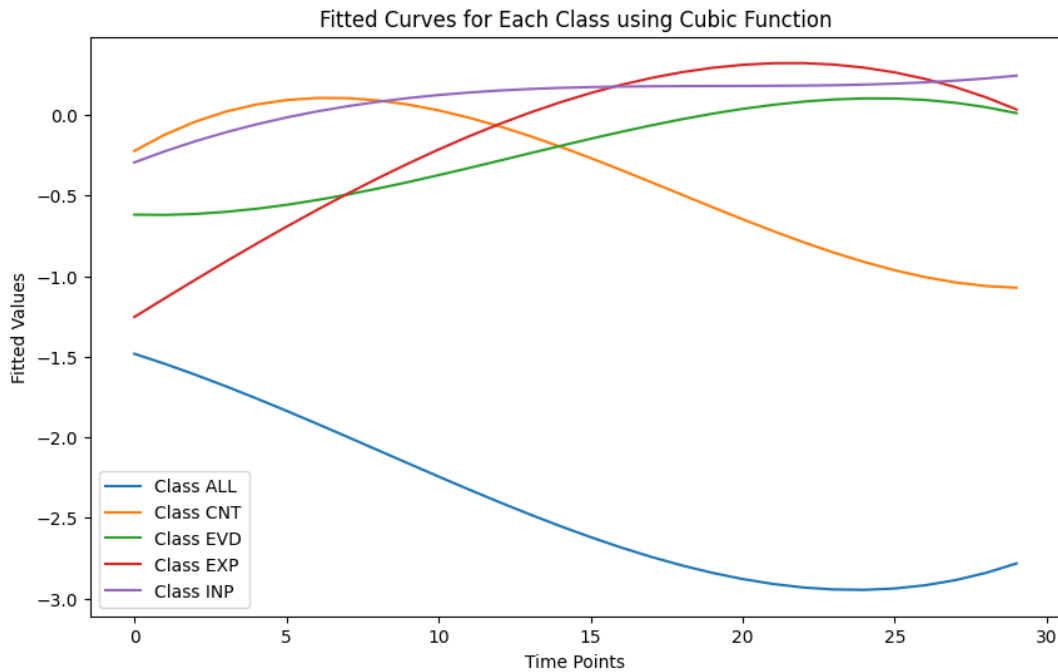
coefficients. This could be expected for EXP and EVD, which are characterized by a rising f_0 movement until the stressed vowel; this movement is potentially followed by a falling movement when segmental material is present. Other features, such as intensity, duration, and position, are good candidates to observe a possible difference. For the CNT-INP pair, the analysis is a little bit more convoluted. Most importantly, as we show, the fitted curves for these two DM functions result in different forms. A possible explanation can be drawn by looking at the boxplots of the four coefficients of both CNT and INP. First, the central tendencies are always very similar for all four coefficients. Second, although INP has a narrower spread, its data is always within the limits of CNT's coefficient distributions.

Table 28 - Significant differences between pairs of DMs by f_0 curve coefficients

DM PAIR		1st coefficient	2nd coefficient	3rd coefficient	4th coefficient
CNT	EXP	✓	✓	✗	✓
CNT	ALL	✗	✗	✓	✓
CNT	INP	✗	✗	✗	✗
CNT	EVD	✓	✓	✗	✓
EXP	ALL	✓	✗	✓	✗
EXP	INP	✓	✓	✗	✓
EXP	EVD	✗	✗	✗	✗
ALL	INP	✗	✗	✓	✓
ALL	EVD	✓	✗	✓	✗
INP	EVD	✓	✓	✗	✓

Figure 38 displays the fitted curves by DM function using a cubic polynomial function. The curves were fitted, always taking 30 f_0 samples regularly spaced along all the DM instances.

Figure 38 - Fitted curves by DM function using a cubic function



The fitted curves are very much in line with initial expectations. They are a good representation of the prototypical curves described in the qualitative analysis. ALL (blue curve) displays a falling f_0 movement from the beginning of the DM, even when there is segmental material before the stressed vowel. Furthermore, ALL curves confirm the observation that this DM function exhibits the lowest f_0 mean compared to the illocutionary unit (COM). The CNT (orange) curve is characterized by the steepest f_0 falling movement along the stressed vowel. This can be seen from time points 5 to 25. However, if CNT has segmental material before the stressed vowel, it is expected to have a rising preparatory movement, making its falling movement more prominent. This can be observed in the data, and the preparation is reflected in the fitted curve from time 0 to 5. From a perceptual standpoint, both the preparation and the different levels of mean f_0 participate in distinguishing between the two DM functions, which may occur in the same position. EVD (green line) and EXP (red line) have similar movements and f_0 levels. However, these two DM functions do not occur in the same position. Moreover, EXP's rising movement is

typically steeper than EVD, which most frequently occurs in the final position. Finally, INP displays the flattest of the forms, as expected. Looking only at the fitted curves of CNT and INP, one can see they are quite distinct. All the same, many INP instances with marked attitude display a rising movement (when there is voiced segmental material) followed by a flat profile (necessary movement) at the stressed vowel and finished by a falling f0 movement (when there is voiced segmental material after the stressed vowel). In such cases, the fitted curve can prove rather insufficient for the distinction, especially considering that, here, the effect of duration is neutralized by the 30-point interpolation. For the distinction of this pair, good candidates are f0 slopes on the stressed vowel and mean standardized duration.

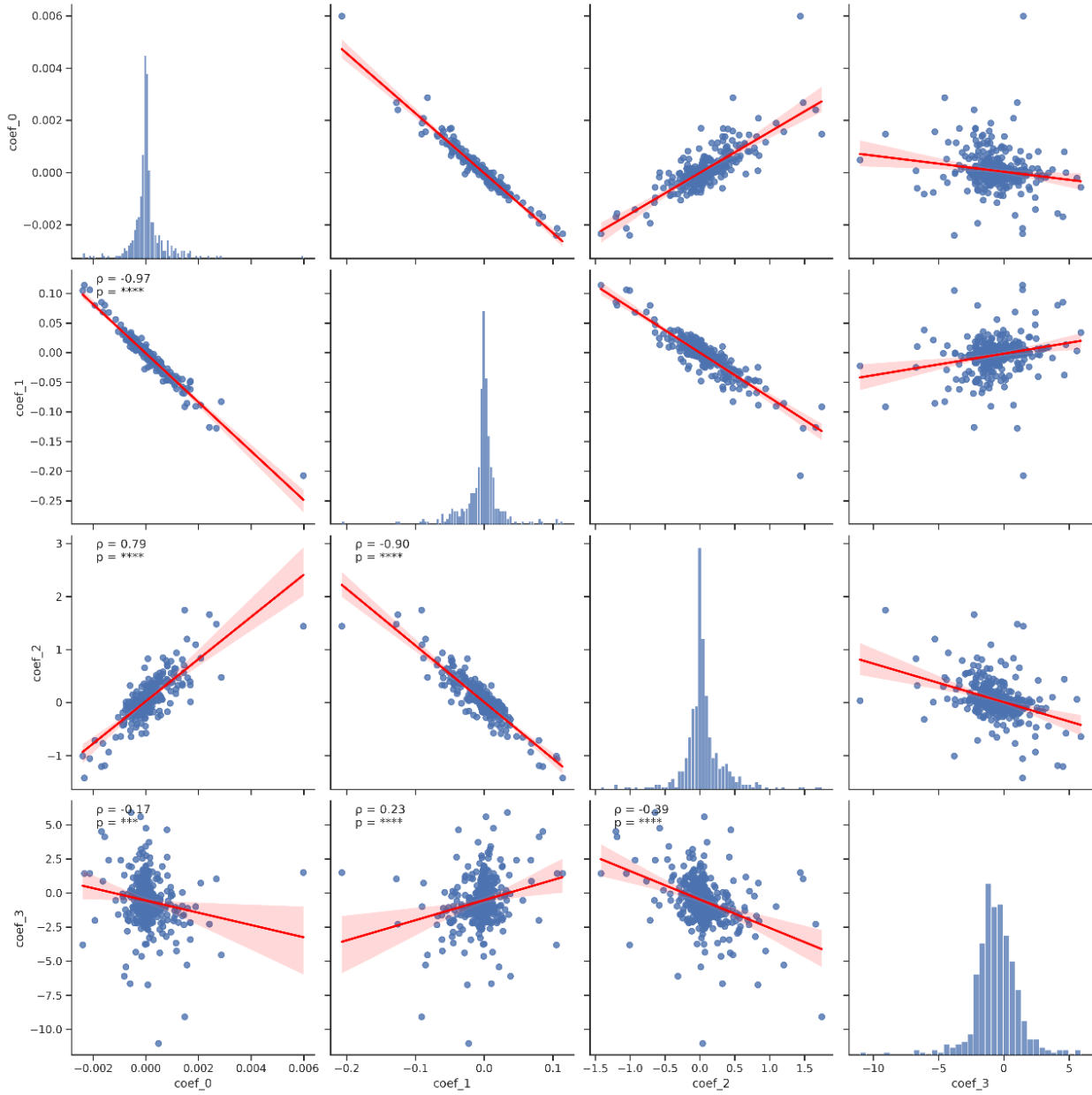
Table 29 shows the summary of statistics, and Figure 39 exhibits the correlation between the coefficients of the fitted curves.

Table 29 - Statistics summary of the features of f0 curve

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
1st coefficient	ALL	0.0002	0.0011	0.000	0.000	-0.002	0.006	0.008	2.260	13.745
	CNT	0.0002	0.0006	0.000	0.000	-0.002	0.003	0.005	1.063	7.891
	EVD	-0.0001	0.0004	0.000	0.000	-0.002	0.001	0.003	-1.187	8.106
	EXP	-0.0001	0.0003	0.000	0.000	-0.001	0.001	0.002	1.077	5.535
	INP	0.0000	0.0005	0.000	0.000	-0.002	0.001	0.004	-1.285	11.266
2nd coefficient	ALL	-0.006	0.041	0.000	-0.004	-0.207	0.105	0.312	-1.614	10.555
	CNT	-0.009	0.027	-0.004	-0.007	-0.128	0.106	0.234	-1.006	9.351
	EVD	0.005	0.019	0.001	0.003	-0.052	0.085	0.137	1.377	9.141
	EXP	0.000	0.015	0.003	0.002	-0.061	0.025	0.086	-1.671	6.352
	INP	-0.003	0.024	-0.001	-0.003	-0.091	0.114	0.205	1.208	12.679
3r	ALL	-0.039	0.384	-0.077	-0.069	-1.007	1.440	2.447	1.318	7.632

Feature	DM	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
	CNT	0.090	0.330	0.024	0.062	-1.056	1.660	2.716	1.321	8.717
	EVD	-0.006	0.253	0.017	0.016	-1.204	0.568	1.772	-1.876	9.789
	EXP	0.113	0.227	0.020	0.074	-0.214	0.941	1.156	1.885	6.245
	INP	0.055	0.348	0.025	0.057	-1.419	1.746	3.164	0.000	14.017
4th coefficient	ALL	-1.364	2.125	-1.080	-1.192	-11.012	2.413	13.424	-1.558	7.947
	CNT	-0.130	1.423	-0.111	-0.195	-4.383	4.758	9.141	0.427	3.797
	EVD	-0.638	1.501	-0.882	-0.684	-6.640	4.521	11.162	-0.079	7.185
	EXP	-1.168	1.265	-0.992	-1.005	-6.720	0.525	7.244	-1.815	7.582
	INP	-0.134	1.923	-0.157	-0.241	-9.064	5.892	14.956	-0.259	9.327

Figure 39 - Correlation between coefficients of the fitted curves



In the next chapter, I will present the different classification models trained and what are the most relevant features for the tasks.

7 CLASSIFICATION MODELS

This section presents the models trained to classify the five DM classes from the prosodic parameters extracted from the corpus. The objective is to show how handcrafted features specifically designed to represent purely prosodic-acoustic features can correctly classify our five DM functions. Before diving into the results of the classification models, the curve fitting procedure is explained. We deal with this problem in this chapter because it specifically considered two factors: the goodness of fit of the curves (as measured by MSE) and the performance of a classification model, taking only the parameters of the best-fitting polynomial coefficients as input. Secondly, the criteria used to evaluate the performances of the classification models are set out. As a starting point, a good model must perform better than a baseline model for a 5-class classification task. The goal is to show that high performance can be met only by using the prosodic features. Then, I move on to comparing different classification techniques. Based on the best-performing technique, I will fine-tune the model and check the most important descriptors (features) for an overall classification model and each DM class against the others.

7.1 CRITERIA FOR ASSESSING DIFFERENT CLASSIFICATION MODELS

Evaluating a good performance in a classification task depends on several factors. This may include the nature of the data, the task's difficulty, the number of observations available, and the application's specific requirements. However, some general guidelines can be drawn from the problem at hand.

I am dealing with a 5-class classification task. The dataset is imbalanced. The CNT function is the majority class and represents as much as 32% of the whole dataset. Gries (2021) recommends that the model should attain a n accuracy score exceeding the percentage of the majority class (so that its performance is not considered chance).

A commonly used benchmark compares the model's accuracy to the baseline accuracy (for the problem at hand, 32% accuracy score). If the model performs significantly better than the baseline, the model is learning valuable patterns from the data. What is considered good can vary depending on the context.

The accuracy score is, of course, one of many metrics to be considered. This is especially true when one is dealing with imbalanced data. Other metrics may come in handy if initial and final CNTs are put in the same class (as I motivatedly do). Precision, recall, F1-score, and the confusion matrix help understand the model's performances. All the same, most of the time, the accuracy score is used as the metric whose best performance is to be pursued.

7.2 F0 CURVE FITTING

Fitting polynomial functions to each observation entails a relevant trade-off. On the one hand, one may increase the degree of the polynomial function and get a lower Mean Squared Error (MSE). On the other hand, one adds details to the curve that are irrelevant to human perception and the model (see the MOMEL stylization process, for example – Hirst & Espesser, 1993). The goal here is not to assess how the curve fitting may adapt to human perception but to find a good compromise between the goodness of fit (how the curves fit the data as measured by the MSE) and the qualitative descriptions for each DM class. As said in the previous chapter, the regression line over the stressed vowel may carry a good deal of information about the DM class. However, a fitted curve can also say something about the general f0 level of the DM instance, where its peaks and valleys occur, and what movements seem to happen throughout the entire DM instance. The fitted curve parameters can ideally bring about distinctive patterns of DM classes outside the stressed vowel. To illustrate this, we can take, as an example, CNT and ALL instances with a pre-stressed syllable. In this region, where CNT is expected to exhibit a preparatory rising movement, ALL will display a falling movement followed by a flat profile. Also, CNT is expected to have a higher f0 level, while ALL will show a lower f0 level. These patterns are not captured by f0 slopes on

the stressed syllable, but the parameters of a polynomial function may ideally account for them.

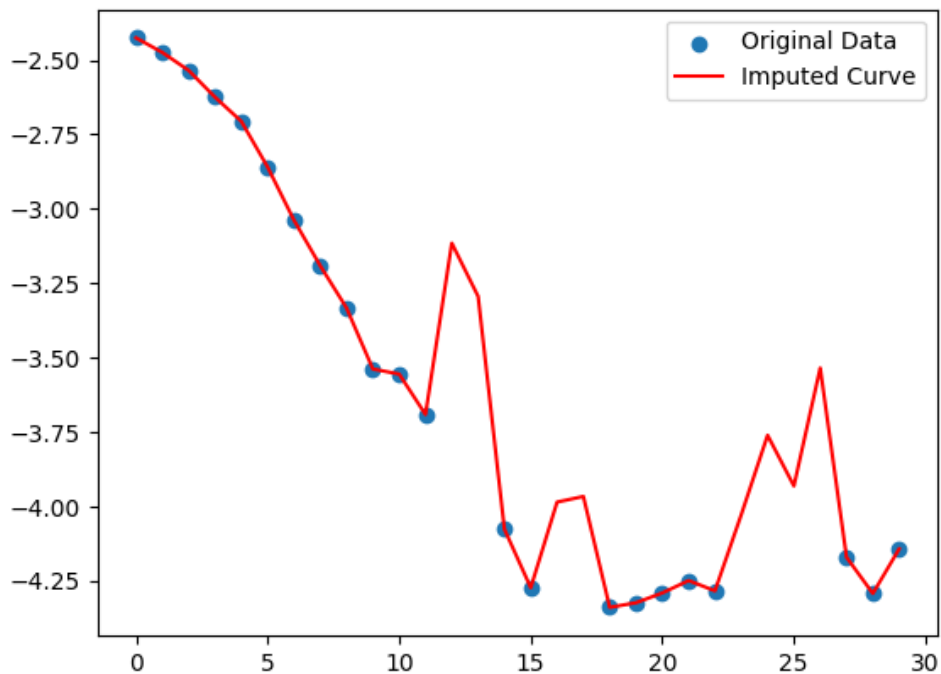
As previously mentioned, I am interested in the f0 curve of the complete DM unit. The normalized f0 points used for the estimations of parameters involving f0 measurements took into consideration only time points that our Voicing Decision model (VD model) classified as voiced. Utilizing only the voiced points is a good heuristic if one wants to avoid potentially deviant estimations that can be caused by fricative sounds, plosive, and sound transition regions. Even if the Viterbi Algorithm was used to smoothen f0 estimations, there might be cases (in the estimation of whole utterances) where all the PDA algorithms may have outputted highly deviant (and erroneous) f0 estimations. However, I have an important consideration to make concerning voicing in speech.

There are two main reasons segmental material is to be produced voiced and unvoiced – which are valid, at least for this research's target language, Brazilian Portuguese. The first one is phonological. In BP (as well as in most languages), voicing is a distinctive trait that allows the language to increase its phoneme inventory. Voicing is thus primarily relevant for the distinction on the morphological and lexical levels. The second one has a broader communicative function. In some types of voice qualities, voicing can be almost absent, such as in voiceless or whispered speech, for which the vocal folds are not actively mobilized (e.g., Laver, 1980). However, I consider that for carrying the pragmatic functions under analysis (that of the DM functions), the more complete the f0 curve is, the better. Of course, there is the possibility that specific voice qualities might be correlated with some DM functions. However, this aspect needs to be further investigated, and it is left out of the scope of this research.

That being said, I needed to close the gaps between voiced and unvoiced segments. Three possible paths were envisaged. The first one filled in the gaps with the averages between the edges of the voiced and unvoiced regions. However, this would result in straight curves, which would be especially problematic when unvoiced segments occurred in regions with falling and rising profiles (Mixdorff & Niebuhr, 2013). A second possibility would be to use some imputation

technique. I used the SoftImpute from the fancyimpute (Rubinsteyn & Feldman, 2016) package for Python and Sci-Kit Learn (Pedregosa et al., 2011) Iterative Imputer to test this possibility. The SoftImpute technique completes a matrix through iterative soft thresholding of Singular Value Decompositions (SVD). This algorithm was inspired by the softImpute R package, which is based on Spectral Regularization Algorithms for Learning Large Incomplete Matrices (Hastie et al., 2014). In turn, the Iterative Imputer utilizes round-robin linear regression that models features with missing values as a function of other features. In both cases, the data matrix is split by class (typically with similar shapes), each timeframe is assumed to be a feature and missing points are imputed as a function of the other data in the matrix. The more variation the data presents, the more the imputed values will be negatively affected. Figure 40 and Figure 41 show the results of these imputers for two different audio files:

Figure 40 - SoftImpute (left) and Iterative Imputer (right) results for file bfamcv07_114



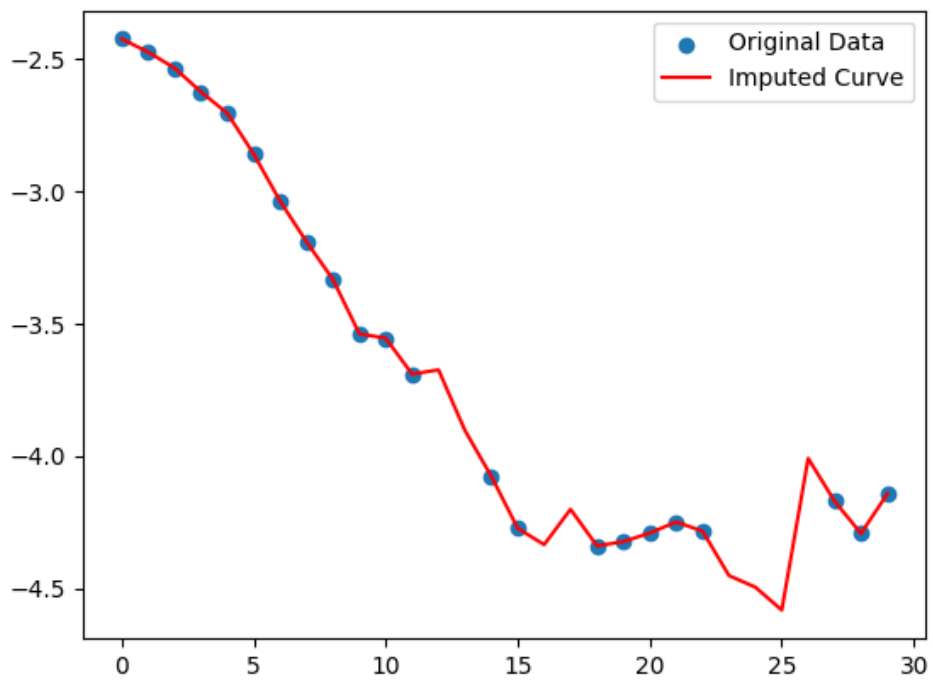
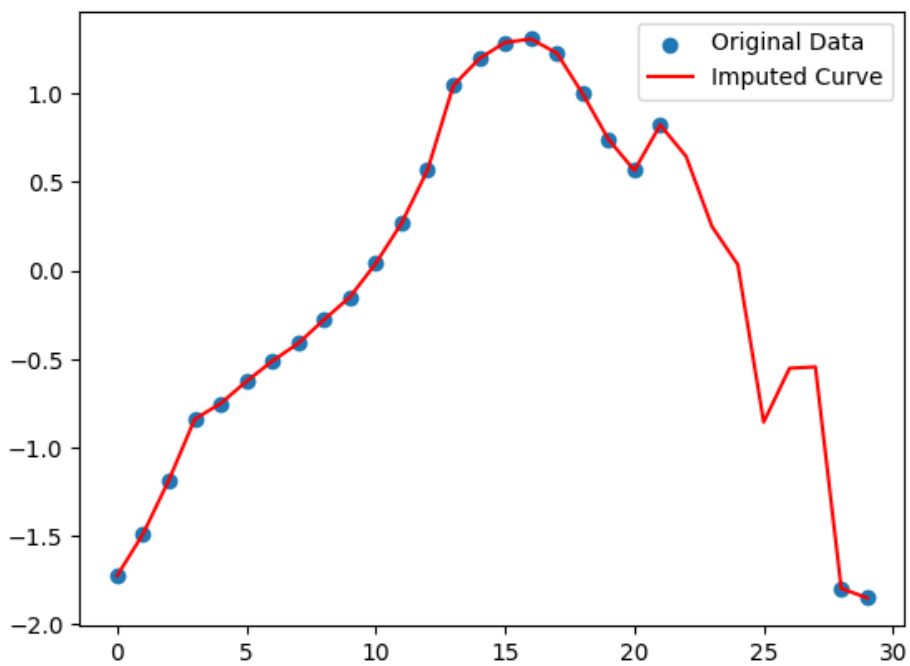
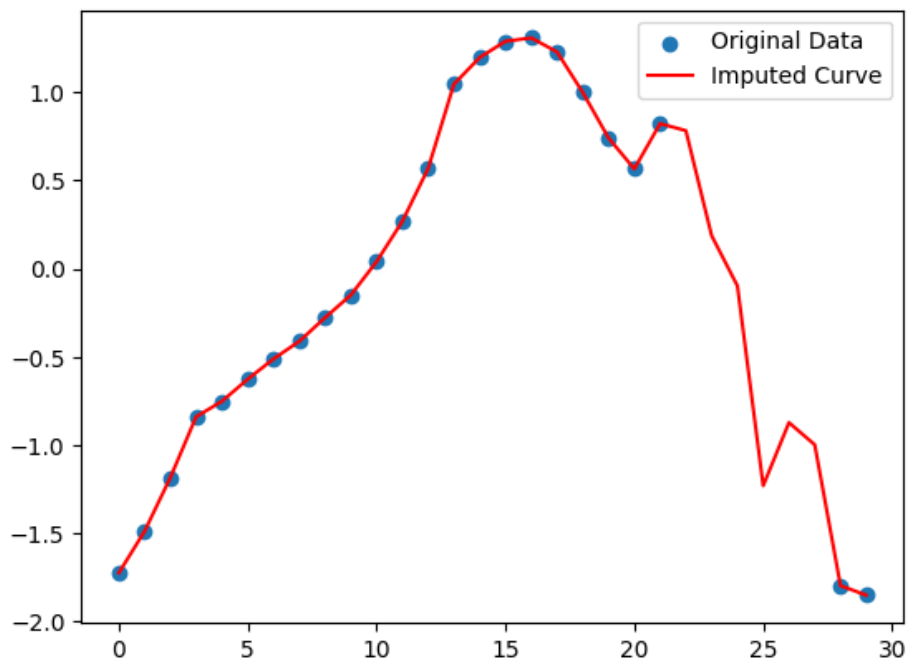


Figure 41 - SoflImpute (left) and Iterative Imputer (right) results for file *bfamcv22_127*





As we can observe, SoftImpute and IterativeImputer work reasonably well for file bfamcv22_127. Here, a simpler polynomial function would easily sweep the noise away, leaving the features of the curve that are relevant for the work. Nonetheless, both imputers create inadmissible noise for the first DM instance (bfamcv07_114 in Figure 40, right). The polynomial curve would certainly be leveled upwards to account for the noise created, at least with SoftImpute. There is a plethora of imputing methods available. Testing their adequacy to the data would require a good deal of work and ground truth f_0 estimations on a 5-class DM dataset especially labelled for this purpose. Since I do not dispose of such data, another heuristic was adopted.

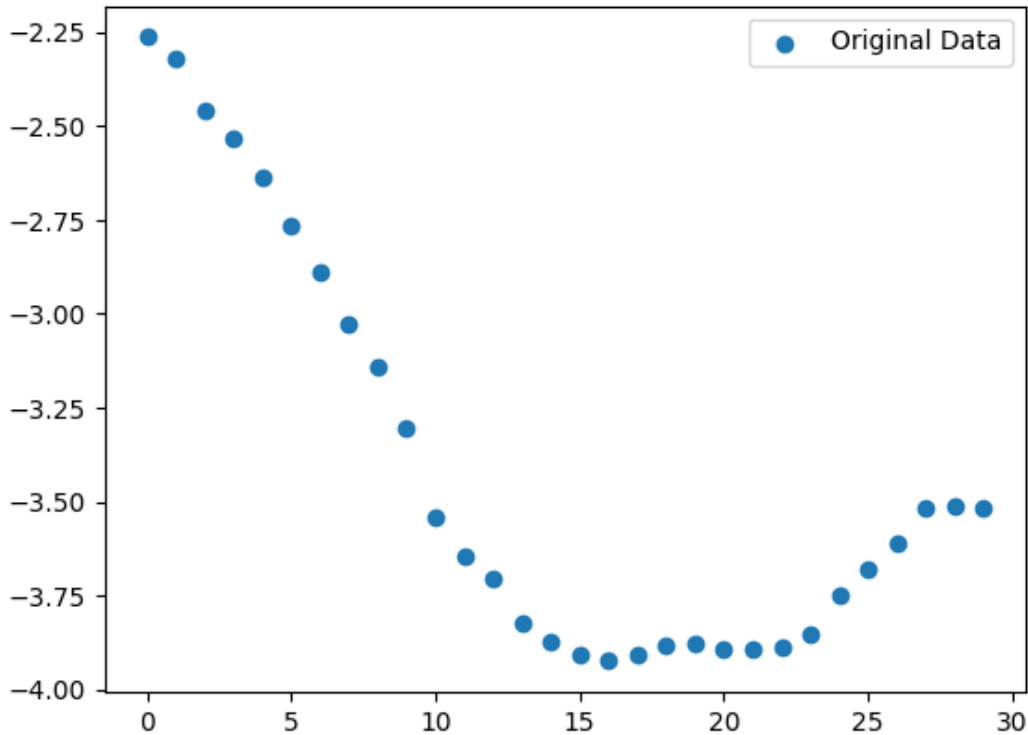
Many PDA algorithms output f_0 estimations for all sampled timepoints. Instead of returning missing values for timepoints that were judged unvoiced, these algorithms do their “best estimations” and output both an f_0 value and a voicing probability. The voicing probability thresholding is not done by the algorithm. It is up to the user to find out what the best threshold is. This can be highly problematic if the user does not have a ground truth voicing decision with respect to which they could minimize the error. But for the

purposes of this work, it can be a good idea to have the estimations of these continuous algorithms through the Viterbi Algorithm such that their estimates result in a continuous curve. Of course, this does not safeguard us against some level of error (especially where all PDA are wrong), but this heuristic can result in curves closer to the reality. At this point, one could argue that (a) simply passing the existing f_0 points to the curve fitting algorithm or that (b) doing a simple linear interpolation between the points of gapping regions would be safer than using any imputation technique or the heuristic adopted. To (a), I can respond that the curve fitting algorithm used does not accept missing values. The data would need imputation before fitting anyway and discarding observations with no missing values would potentially result in an empty matrix. To (b), there is the problem that nothing could guarantee beforehand that there would be voiced points on both borders of regions with missing values. This would rule out the linear interpolation. For instance, in Figure 41, if all f_0 estimations after timepoint 15 were missing, no reasonable interpolation would be possible.

The six PDAs whose estimations were available for all timepoints were employed. They are namely PEFAC (Gonzalez & Brookes, 2014), RAPT (Talkin & Kleijn, 1995), SWIPE (Camacho, 2007), SWIPEP (Camacho, 2007), SRH (Degottex et al., 2014), and YIN (de Cheveigné & Kawahara, 2002). Their output was interpolated for 30 equally spaced timepoints. This was done for two reasons: one, because the curve-fitting algorithm always needs the same number of points; and two, because it neutralizes the curve's durational differences. Figure 42 shows the result of this procedure for file *bfamcv07_114*, which presented more issues when imputation methods were applied:

Figure 42 - F_0 curve of six selected PDAs smoothed by Viterbi

Algorithm



With the continuous data, a curve-fitting algorithm was used to test six potential polynomial functions, ranging from the linear (two coefficients/parameters) to the sextic function (with 7 coefficients/parameters). Furthermore, six different classification models were trained, taking only the fitted curves' parameters as input. These models aimed to assess to what degree a model based solely on f_0 curve features can correctly predict the five DM classes. The models were evaluated on a stratified 5-fold cross-validation set. Only one classification technique was used, the Linear Discriminant Analysis, which exhibited some of the best accuracy scores for the tasks carried out in this chapter, as shown further ahead. Furthermore, the number of observations in each class was balanced. This was done because the CNT class has almost double the size of other classes that tend to occur in fixed positions (CNT can occur in initial and final positions). Table 30 shows the mean MSE value, the standard deviation of MSE values, and the mean accuracy score of the 5-fold cross-validation task:

Table 30 - Classification performance based on f_0 curve coefficients

Poly function	Mean MSE	STD MSE	Mean Accuracy score
Linear	0.67	2.65	0.48
Quadratic	0.48	1.98	0.51
Cubic	0.38	1.54	0.51
Quartic	0.29	1.04	0.50
Quintic	0.22	0.77	0.52
Sextic	0.18	0.64	0.50

Performance above 0.5 is achieved using all polynomial function coefficients of 2 or more degrees. The quadratic function would already exhibit satisfactory results. However, one of the goals of this work is to find the prototypical curves of each DM class. By inspecting the results of the fitted curves against the actual f_0 data, we can see that the quadratic function (Figure 43) will oversimplify the data when we compare it against the cubic function (Figure 44). Figure 45 displays the fitted curve for the quartic function.

Figure 43 - Fitted curve vs original data of the quadratic function

(Audio file bfamcv11_2)

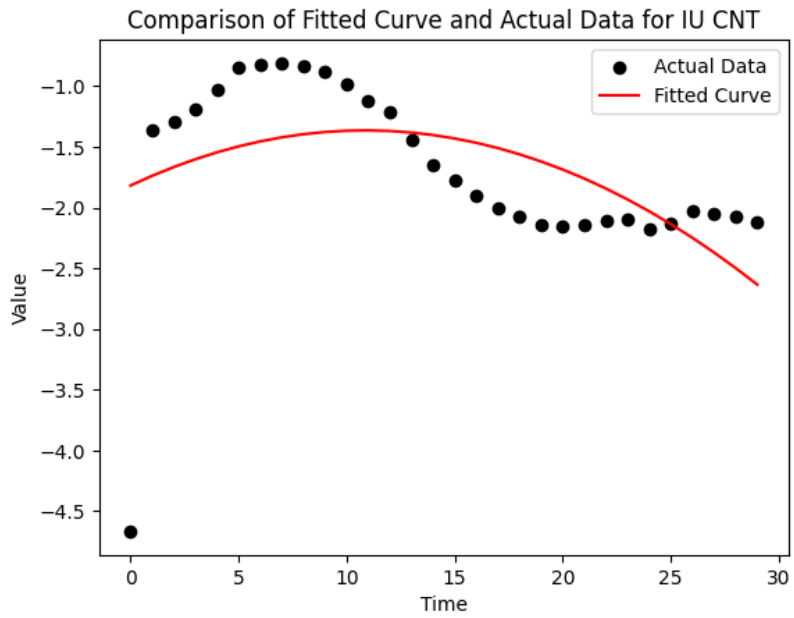


Figure 44 - Fitted curve vs original data of the cubic function (Audio file bfamcv11_2)

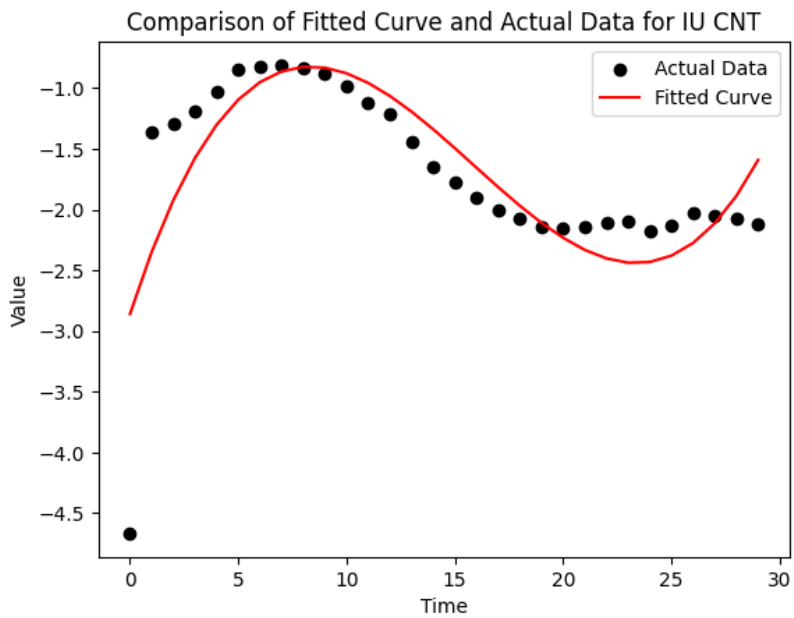
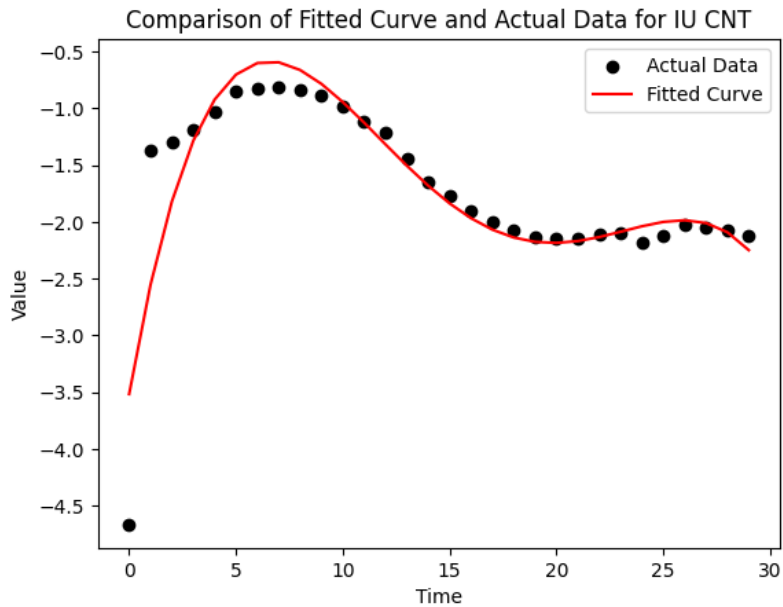


Figure 45 - Fitted curve vs original data of the quartic function (Audio

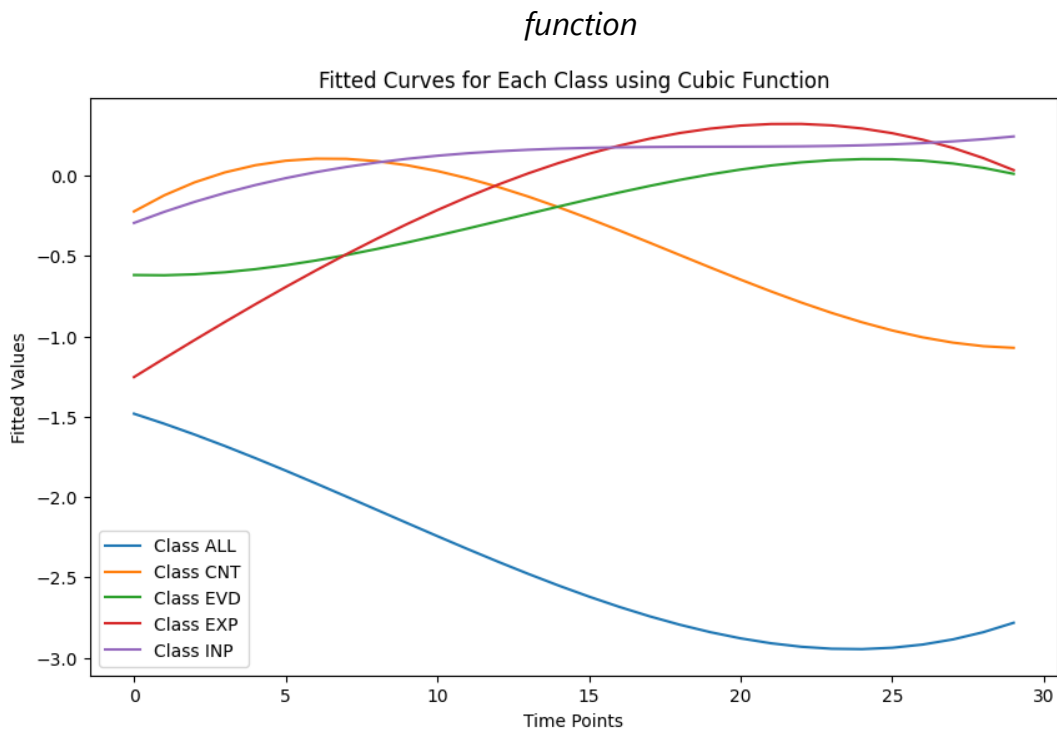
file bfamcv11_2)



In the examples above, the quartic function fits the original curve much better. However, if we check the accuracy score of the model that uses its parameters, we can observe that the quartic function leads to the second worst results among the tested functions. To avoid creating more parameters and, at the same time, oversimplifying the data, it was decided to use the coefficients of the cubic function.

To create a visualization of the prototypical curves, the data was split into the five DM classes and the curves were fitted to the resulting matrices using the cubic function. Figure 46 shows the resulting curves.

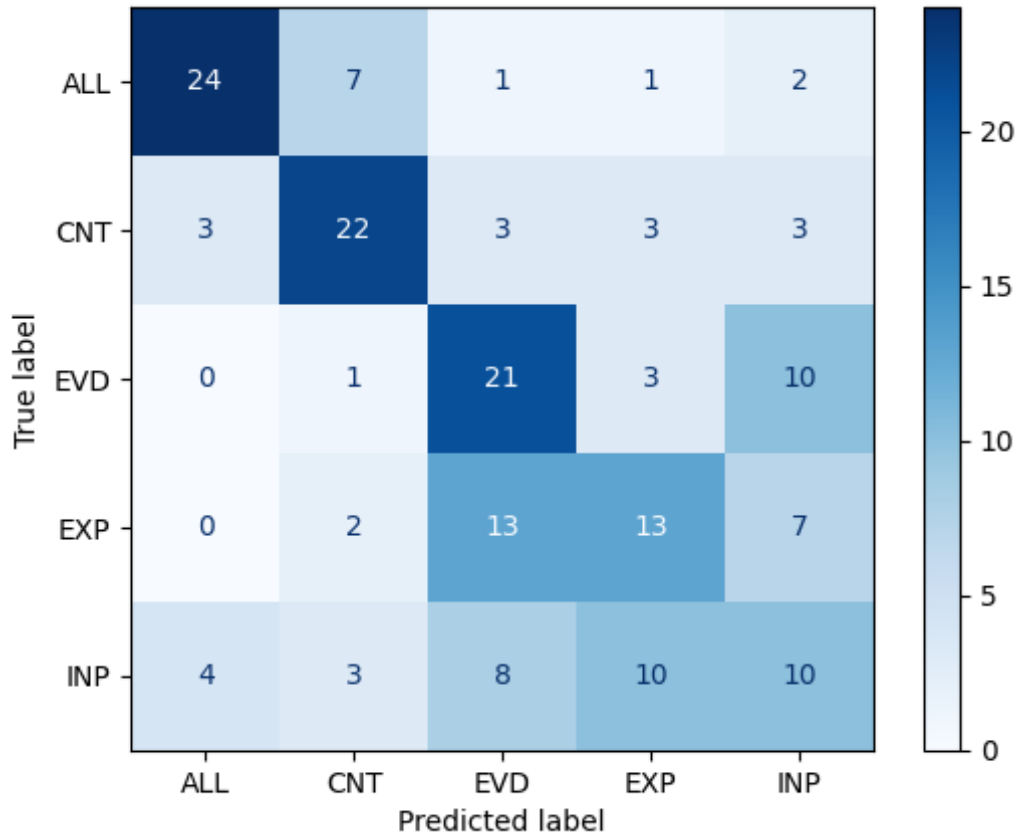
Figure 46 - Prototypical f_0 curves of each DM class using the cubic



We can now look at the confusion matrix resulting from the LDA classification model taking as input only the coefficients of the cubic function in Figure 47. Note that this model has no information about the position of DM instances.

Figure 47 - Confusion matrix for an LDA model using coefficients of the

cubic function



As we can see, the best-performing class is ALL. As already mentioned in the previous chapter, this class has, in general, the most distinctive curve in terms of form and level. It starts to fall from the beginning, and it has the lowest f₀ level of all forms. If we consider only f₀ level, we can see that CNT, EXP and EVD have almost the same level. This can be checked on the descriptive statistics of mean f₀ in Table 23 in the previous chapter. However, CNT has the most distinctive form, and this is reflected in the classification accuracy score of this class, which is the second best performing. In third place, we have the EVD class. This class is generally not mixed with others but INP. This is because INP can have a slightly rising f₀ curve, just as EVD. The DM's position and other parameters will play a crucial role in class distinction here. The distinction of EXP and INP on the base of f₀ curves also represent a crucial problem. INP has a flat profile on the stressed vowel but when it has voiced segmental material before, it may display a rising profile. This will be translated into a curve similar to EXP. Many EXP and INP

instances can thus be easily confounded when taken in isolation and without phonemic information. In the next section, I put the f0 curve parameters together with the other features to analyze the performance of different baseline models.

7.3 CHOOSING A CLASSIFICATION TECHNIQUE

The last section dealt with models that only took into account f0 features. From here, I begin by selecting a classification technique using all prosodic features. The baseline classifiers presented here will serve for the selection of a technique that will be used in the next steps of this work. At this point, the models are trained using the whole dataset. As a reminder, the dataset presents some imbalancedness, especially for the CNT class, which has as much as double the observations as the other classes (but can be found in two positions). The amount of data of the other classes ranges between 68 and 80 observations (see Table 18).

Table 31 - Number of observations and proportions per DM class

Class	Observations	Proportion
ALL	68	0.16
CNT	139	0.32
EVD	75	0.17
EXP	69	0.16
INP	80	0.19

For training the models, a simple stratified train/evaluation split with the ratio (0.75/0.25) was adopted. Since the number of observations is limited, training was done on the whole train subset. 12 classification techniques were evaluated: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Naïve Bayes (NBC), Random Forest (RFC), Gradient Boosting (GBC), Bagging (BAG), AdaBoost (ADA), Decision Tree (DTC), Support Vector Machine (SVC), Logistic Regression (LGC), and Multi-layer Perceptron (MLP). A basic description of these algorithms, as well as their pros and

cons are summarized below, based on Géron (2022) and on the documentation of Scikit-Learn (Pedregosa et al., 2011):

LDA and QDA. Both methods try to find a combination of features that characterizes or separate two or more classes. LDA uses a linear whereas QDA uses a quadratic decision surface. Due to their intrinsic multiclass nature, quickly computed closed-form solutions, shown practical effectiveness, and lack of tuning hyperparameters, both approaches can be very effective and easy to handle.

KNN. The K-Nearest Neighbors algorithm stores train data in vectors and then compares new data to stored data based on a selected metric – typically the Euclidean distance. KNN is simple to implement and handles multi-class tasks well with enough representative observations. On the other hand, its hyperparameter tuning can be more complicated (need to find k best value and best comparison metric), and computation is costly (it does not handle large datasets with too many features well).

NBC. The Naïve Bayes is a simple algorithm that can effectively handle large datasets. It works especially well for text classification tasks like sentiment analysis and spam filtering. Naïve Bayes simplifies computation and reduces overfitting by assuming that characteristics are conditionally independent of the classes. For this reason, the algorithm performs effectively even with a small amount of training data. Because Naïve Bayes is resistant to irrelevant features, it is appropriate for high-dimensional data²⁵. Naïve Bayes assumes that features are independent, which is not the case in most real-case circumstances. In cases where the independence assumption is broken, this may result in less-than-ideal performances. Also, it may not capture intricate feature interactions. Compared to

²⁵ High-dimensional data are defined as data in which the number of features (parameters) is close to or larger than the number of observations.

more complex algorithms, this may reduce accuracy, particularly when working with highly correlated features. For Naïve Bayes to accurately estimate the class probabilities, adequate training data must be available.

RFC. Random Forest is an ensemble learning technique²⁶ that combines several decision trees to provide predictions. It is appropriate for various classification tasks due to its strong accuracy and resilience against overfitting. Random Forest can handle numerical and categorical features without requiring a lot of data preprocessing. Additionally, it has good handling power for outliers and missing values. By offering feature importance measures, Random Forest helps users comprehend the relative significance of various aspects throughout the classification process. This can help with feature selection and prediction interpretation. On the other hand, this algorithm can be computationally costly, particularly when working with big datasets or an ensemble of several trees. Training and evaluation times could go up a lot. Predictions can be biased as a result of its tendency to favor the majority class, and RFC may not work well for imbalanced data. When compared to individual decision trees, Random Forest can be challenging to interpret since its ensemble nature makes it difficult to comprehend the underlying decision-making process.

GBC. Using a series of weak learners²⁷, usually decision trees, the Gradient Boost Classifier is an ensemble learning algorithm that generally exhibits a powerful predictive capacity. It has a reputation for being very accurate and capable of handling complicated datasets. Gradient Boosting is flexible for handling a variety of data types (numerical and categorical features). Due

²⁶ A technique that creates an ensemble of submodels and/or multiple subsets of the data under the hood. The base estimators can vary depending on the algorithm. Random Forest, for instance, always use Decision Trees.

²⁷ In ensemble learning, weak learners are submodels that perform better than random guesses, whereas strong learners exhibit good accuracy scores.

to its ability to give minority classes greater weights, it also performs well on datasets that are imbalanced. By offering feature importance metrics, gradient boosting enables users to comprehend the relative significance of several characteristics throughout the classification process. This can help with model interpretation and feature selection. Gradient boosting can be costly and time-consuming in terms of computation, particularly when working with big datasets or a lot of weak learners. To avoid overfitting and attain peak performance, hyperparameters like learning rate and tree depth must be optimized. Because gradient boosting can easily pick up noise and outliers, it may not work well on noisy or sparse datasets. In these situations, data pretreatment and feature selection are essential to enhancing performance.

BAG. The Bagging Classifier is also an ensemble learning technique that generates predictions by combining several base estimators (models). Different from Random Forest, the base estimator must be chosen by the user. By lowering variance and overfitting, it helps raise the model's overall stability and accuracy. Bagging can handle both numerical and categorical features. Additionally, it has good handling power for outliers and missing values. Large datasets can benefit from bagging because it is parallelizable and computationally efficient. Also, it can offer class probability estimates, enabling more complex predictions. Bagging does not work well with imbalanced datasets, and its predictions are biased due to its tendency to favor the majority class. Individual base classifiers are easier to understand than bagged data. The algorithm's ensemble nature makes it difficult to comprehend the underlying decision-making process. High-dimensional data might not be good candidates for bagging. It may result in overfitting and in increased computational complexity.

ADA. AdaBoost Classifier is another ensemble technique that joins several weak learners, usually decision trees, to produce a powerful predictive model. The base estimator must also be tweaked by the user. It is reputed for being very accurate and capable of handling complicated datasets. AdaBoost is

especially good at managing imbalanced datasets because it lets the model concentrate on the minority class. It can handle both numerical and categorical features. It also offers feature importance measurements, making interpretation easier. AdaBoost is susceptible to noisy data and outliers, which could result in overfitting. This technique can be costly and time-consuming in terms of computation, particularly when working with big datasets or a large number of weak learners. If the weak learners are overly complex or prone to overfitting, AdaBoost could not work effectively. Selecting suitable weak learners and fine-tuning hyperparameters are crucial in avoiding overfitting and attaining peak efficiency.

DTC. The Decision Tree Classifier is an easy-to-interpret technique that efficiently manages numerical and categorical features. It is appropriate for complicated datasets because it can manage non-linear interactions between features and the target variable. Decision trees do not need a lot of data preprocessing to handle outliers and missing values. Overfitting is a common problem with decision trees, particularly when the tree grows too intricate or deep. Small changes in the data may cause decision trees to react differently, resulting in various tree architectures and possibly different predictions. Due to their propensity to favor the majority class and inability to reliably anticipate the minority class, decision trees may not perform well on imbalanced data.

SVC. Support Vector Machines is a technique that maximizes the margin between classes to identify the ideal hyperplane for dividing the data. SVM works well when there are more features than observations (high-dimensional data). By utilizing kernel functions, it may also manage non-linear interactions between features. Some of SVM's drawbacks include its sensitivity to the selection of the kernel function and hyperparameters. Computational costs may also be high, particularly for huge datasets.

LRC. Logistic Regression is an algorithm for binary classification problems. It uses a logistic function to model the relationship

between the features and the probability of pertaining to a specific class. Logistic Regression excels in handling huge datasets. The findings are also interpretable because the coefficients may be utilized to comprehend how each feature affects the estimated probability. On the other hand, the algorithm assumes a linear relationship between the target log-odds and the features. If the relation is non-linear, it might not function well. Moreover, it is susceptible to outliers when there is multicollinearity among features. These problems can be lessened with regularization techniques like L1 or L2 regularization.

MLP. A simple Feed-forward Neural Network Classifier consisting of at least input, hidden, and output layers that process data in a forward direction. It can model complex non-linear relationships and is widely used in image and speech recognition applications. However, it requires a large amount of data to train effectively and can be prone to overfitting without proper regularization.

SVM and Logistic regression models are classifiers specialized in binomial classifications. In the Sci-kit learn package, if these models receive multiclass data, they adopt a One-Vs-Rest classification strategy by default. This strategy is characterized by building multiple models that make binomial classifications sequentially.

Table 32, below, exhibits the classification report of each classifier. The classification report contains metrics of precision, recall, f1-score, micro-accuracy/f1-score, macro-accuracy/f1-score, and average accuracy/f1-score. These metrics are presented below:

Equation 8 - Precision

$$Precision = \frac{TP}{TP + FP}$$

Where,

TP is the number of True Positives; and

FP is the number of False Positives.

Recall measures, of all CNT in the dataset, how many of them the model predicted as CNT.

Equation 9 - Recall

$$Recall = \frac{TP}{TP + FN}$$

Where,

TP is the number of True Positives; and

FN is the number of False Negatives.

The f1-score is the harmonic mean of precision and recall, calculated as

Equation 10 - F1-score

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

In multi-class classification, f1-scores are calculated for each class in a One-vs-Rest (OvR) approach instead of a single overall f1-score, as for binary classification. The micro-accuracy and the micro f1-score are the proportions of correctly classified observations out of all observations. The accuracy score is simply the number of correct predictions divided by the number of observations of each class in the support (number of evaluated instances). The number is the same for the global accuracy and the f1-score, presented in the merged line. In their turn, the macro averages are the arithmetic mean either of the accuracy score or the f1-score. Finally, the weighted averages are calculated by taking the

means of all classes weighted by each class's support. The focus is on the global accuracies (accuracy in the merged lines).

Table 32 - Classification report of different classifiers

MODEL / TECHNIQUE	CLASS	PRECISION	RECALL	F1-SCORE	SUPPORT
LDA	ALL	0.86	0.71	0.77	17
	CNT	0.79	0.63	0.70	35
	EVD	0.70	0.84	0.76	19
	EXP	0.62	0.76	0.68	17
	INP	0.59	0.65	0.62	20
	accuracy	0.70			
	macro avg	0.71	0.72	0.71	108
	weighted avg	0.72	0.70	0.70	108
QDA	ALL	0.53	0.53	0.53	17
	CNT	0.43	0.46	0.44	35
	EVD	0.86	0.63	0.73	19
	EXP	0.56	0.53	0.55	17
	INP	0.46	0.55	0.50	20
	accuracy	0.53			
	macro avg	0.57	0.54	0.55	108
	weighted avg	0.55	0.53	0.53	108
KNN	ALL	0.67	0.59	0.63	17
	CNT	0.70	0.66	0.68	35
	EVD	0.65	0.68	0.67	19
	EXP	0.48	0.65	0.55	17
	INP	0.59	0.50	0.54	20
	accuracy	0.62			
	macro avg	0.62	0.62	0.61	108
	weighted avg	0.63	0.62	0.62	108
NBC	ALL	0.56	0.29	0.38	17
	CNT	0.52	0.46	0.48	35
	EVD	0.68	0.79	0.73	19
	EXP	0.50	0.71	0.59	17
	INP	0.45	0.50	0.48	20
	accuracy	0.54			
	macro avg	0.54	0.55	0.53	108

MODEL / TECHNIQUE	CLASS	PRECISION	RECALL	F1-SCORE	SUPPORT
	weighted avg	0.54	0.54	0.53	108
RFC	ALL	0.83	0.59	0.69	17
	CNT	0.67	0.83	0.74	35
	EVD	0.75	0.79	0.77	19
	EXP	0.47	0.53	0.50	17
	INP	0.50	0.35	0.41	20
	accuracy	0.65			
	macro avg	0.65	0.62	0.62	108
	weighted avg	0.65	0.65	0.64	108
GBC	ALL	0.92	0.71	0.80	17
	CNT	0.69	0.83	0.75	35
	EVD	0.72	0.68	0.70	19
	EXP	0.53	0.59	0.56	17
	INP	0.50	0.40	0.44	20
	accuracy	0.67			
	macro avg	0.67	0.64	0.65	108
	weighted avg	0.67	0.67	0.66	108
BAG	ALL	0.91	0.59	0.71	17
	CNT	0.74	0.83	0.78	35
	EVD	0.81	0.89	0.85	19
	EXP	0.61	0.82	0.70	17
	INP	0.64	0.45	0.53	20
	accuracy	0.73			
	macro avg	0.74	0.72	0.72	108
	weighted avg	0.74	0.73	0.72	108
ADA	ALL	0.00	0.00	0.00	17
	CNT	0.64	0.80	0.71	35
	EVD	0.67	0.95	0.78	19
	EXP	0.47	0.88	0.61	17
	INP	0.40	0.10	0.16	20
	accuracy	0.58			
	macro avg	0.43	0.55	0.45	108
	weighted avg	0.47	0.58	0.49	108
DTC	ALL	0.65	0.65	0.65	17
	CNT	0.74	0.66	0.70	35
	EVD	0.74	0.74	0.74	19
	EXP	0.62	0.76	0.68	17

MODEL / TECHNIQUE	CLASS	PRECISION	RECALL	F1-SCORE	SUPPORT
	INP	0.45	0.45	0.45	20
	accuracy	0.65			
	macro avg	0.64	0.65	0.64	108
	weighted avg	0.65	0.65	0.65	108
SVC	ALL	0.88	0.41	0.56	17
	CNT	0.68	0.77	0.72	35
	EVD	0.60	0.63	0.62	19
	EXP	0.47	0.47	0.47	17
	INP	0.43	0.50	0.47	20
	accuracy	0.59			
	macro avg	0.61	0.56	0.57	108
	weighted avg	0.62	0.59	0.59	108
LGC	ALL	0.83	0.59	0.69	17
	CNT	0.74	0.74	0.74	35
	EVD	0.70	0.74	0.72	19
	EXP	0.50	0.53	0.51	17
	INP	0.52	0.60	0.56	20
	accuracy	0.66			
	macro avg	0.66	0.64	0.64	108
	weighted avg	0.67	0.66	0.66	108
MLP	ALL	0.90	0.53	0.67	17
	CNT	0.72	0.74	0.73	35
	EVD	0.76	0.84	0.80	19
	EXP	0.50	0.53	0.51	17
	INP	0.52	0.60	0.56	20
	accuracy	0.67			
	macro avg	0.68	0.65	0.65	108
	weighted avg	0.69	0.67	0.67	108

The models exhibiting the best overall performances are in descending order the Bagging classifier (0.73), the LDA model (0.70) and the MLP model (0.67), as measured by the micro-averaged f1-score. Because it is easier to hyperparameter tune while showing good results, the LDA model was selected for the next steps of this work: balancing the data, conducting a feature selection, and training and evaluating an overall model with more robust evaluation techniques (stratified k-fold and

Leave-One-Out cross-validation sets).

7.4 FEATURE SELECTION WITH LEAPS AND BOUNDS

Feature selection is the deployment of algorithms to reduce the dimensionality of data and improve model performance. There are many reasons why a feature selection algorithm can be used. They can make the model simpler and faster to run. They may improve performance by removing irrelevant features and thus making the data more compatible with the modelling technique. They can be used to avoid the curse of dimensionality²⁸. More importantly for this research, feature selection can be used to show which of them are more mobilized to predict target classes, thus helping understand the model.

To carry out this task, the Leaps and Bounds algorithm (Furnival & Wilson, 1974, implemented in R by Lumley & Miller, 2004) was used. This algorithm was used in Gobbo (2019) for a 3-class classification model with good results. The primary purpose of the Leaps and Bounds algorithm is to explore subsets of features so as to identify the best combination for building a regression model. The algorithm is particularly useful when the number of potential features is high, thus making computationally expensive or impractical to test all possible combinations. 30 features would result in 2^{30} different models to be evaluated, which is not a reasonable solution.

The Leaps and Bounds algorithm works by making leaps (Forward Selection). It starts with an empty set of predictor features. At each step, it adds the features that results in the highest improvement in model fit until a predetermined stopping criterion is reached. The algorithm continues this procedure until the stopping criterion is met. Then, the algorithm bounds (Backward Elimination). It starts with the full set of predictor features. At each step, it removes the feature that has the least impact on the model fit and continues until the stopping

²⁸ Increasing the number of features in a problem entails exponentially increasing the number of observations for the model to be reliable (Bellman et al., 1957).

criterion is met. The Leaps and Bounds algorithm allow for an efficient exploration of the feature space, providing a good compromise between the forward selection and backward elimination methods. The subset of features that yields the best fit according to a specified criterion (e.g., adjusted R-squared, AIC, BIC) is then picked as the final model. While the Leaps and Bounds algorithm was proposed in the context of linear regression, similar concepts and principles can be extended to other regression techniques. This is the case with the Linear Discriminant Analysis (LDA). In R, the leaps package provides a function called `regsubsets()` that implements the leaps and bounds feature selection method.

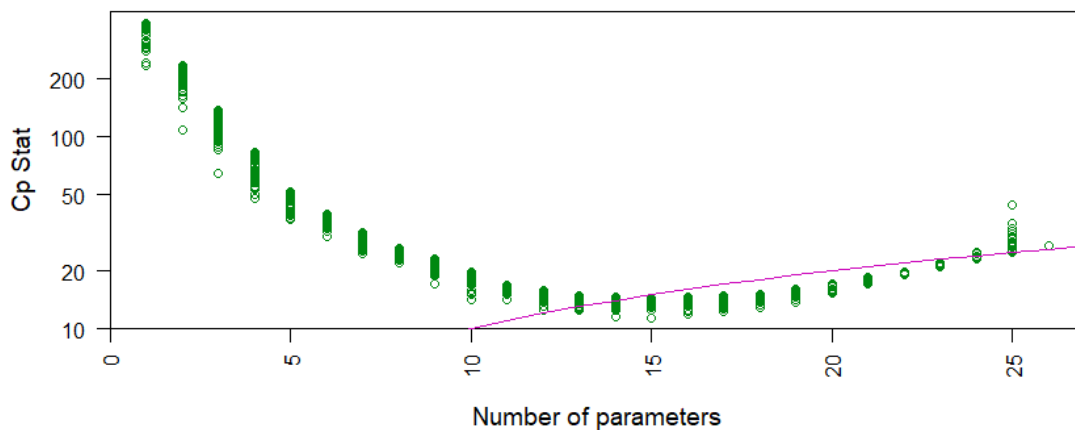
The same procedure followed in Gobbo (2019) was adopted here. However, I evaluate the models on an overall 5-class and a One-Vs-Rest approach. The selection criteria for models of different sizes was the Mean Square Error (MSE) in prediction, as given by Mallows's Cp statistic (Mallows, 1973; 2000). In statistics, Mallows's Cp is used to evaluate the fit of regression models. It is applied in the context of feature selection, where many features are available for predicting some outcome. Smaller Cp values, typically between 0 and 1, indicate that the model is relatively precise.

The script used for this purpose was implemented by Gobbo (2019). It uses the `regsubset()` function from the leaps package. This function finds the best models for each number of parameters. The function was set to use all features available and 431 DM instances and search for a maximum of 60 best combinations chosen from all possible subsets of features. This resulted in 1679 combinations. The number 60 was arbitrarily chosen as large enough without excessive computational cost. This number was tested by Gobbo (2019) with meaningful results.

7.4.1 Global model

I begin with the global model (for the 5 DM classes). Figure 48 below shows Cp values for the 60 best combinations of features as a function of the number of parameters for all the combinations chosen by Leaps and Bounds:

Figure 48 - Estimation of MSE (C_p) for the 60 best combinations of features

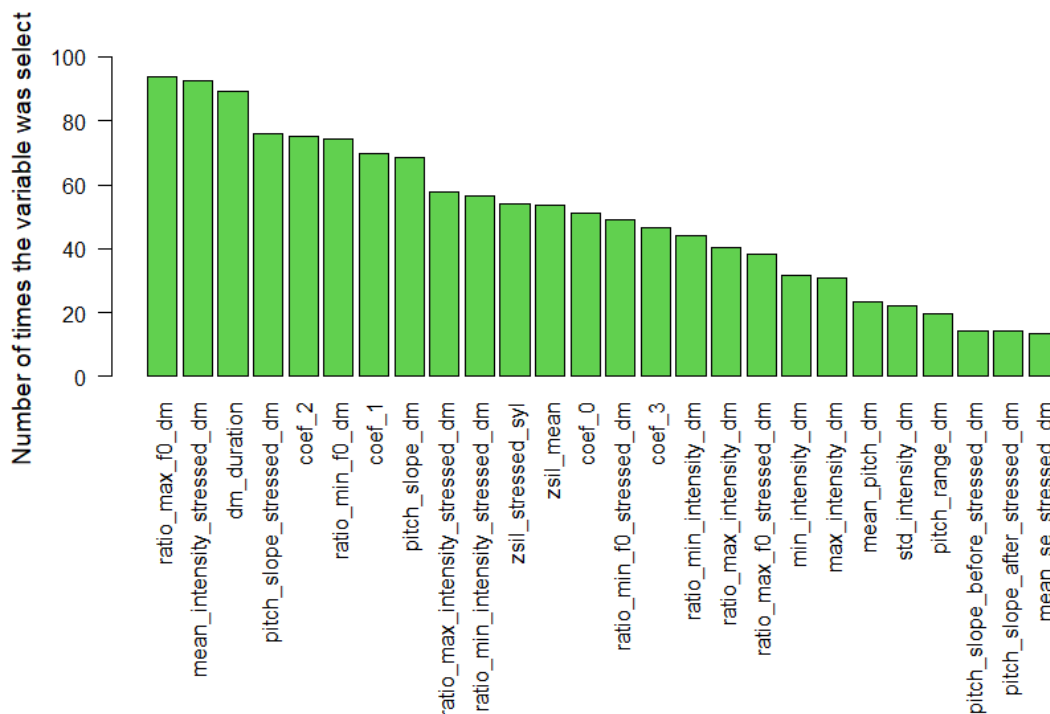


The lowest C_p value suggests a number of 15 features for the best model. The number of times each feature was selected in the 1679 models was calculated. Figure 49 shows the results. We can observe that the 15 most selected parameter were, in descending order:

1. the alignment f_0 peak inside the DM instance
2. mean relative intensity in the stressed vowel
3. relative DM duration,
4. f_0 slope in the stressed vowel
5. 3rd f_0 curve coefficient
6. alignment of f_0 valley inside the DM instance
7. the 2nd f_0 curve coefficient
8. f_0 slope in the whole DM instance
9. Alignment of max intensity with respect to the stressed vowel
10. Alignment of min intensity with respect to the stressed vowel
11. Standardized duration of the stressed syllable

12. 1st coefficient of the f0 curve
13. Alignment of min f0 inside the DM instance
14. 4th coefficient of the f0 curve
15. Alignment of min intensity inside the DM instance

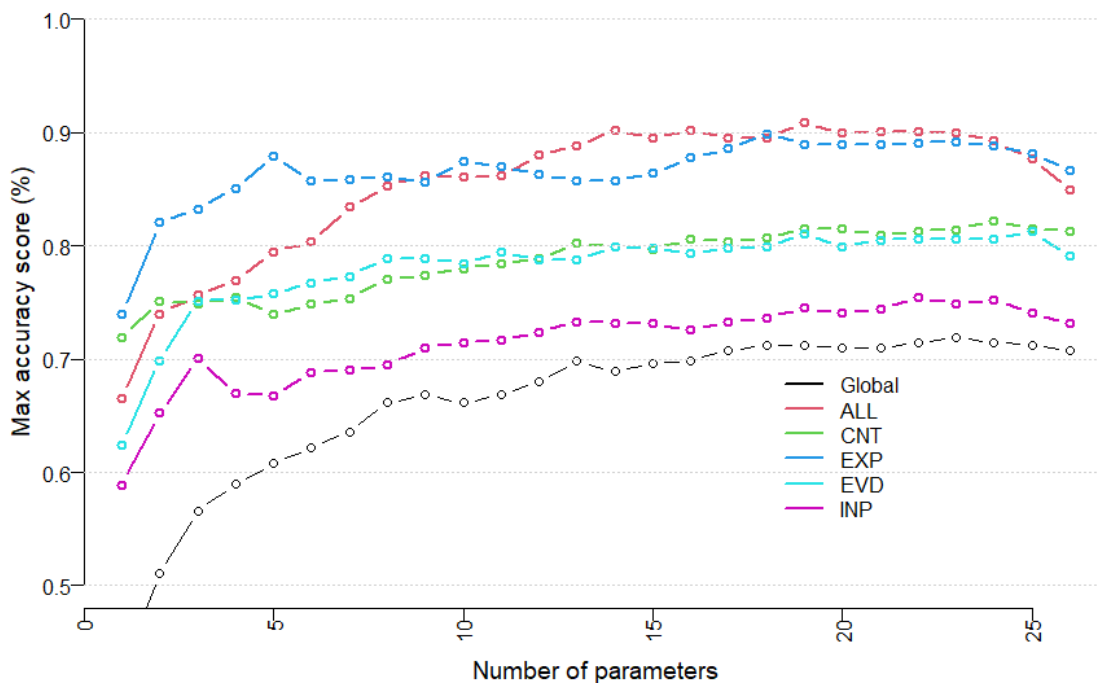
Figure 49 - Number of times each feature was selected among best combinations (%)



Each of the 1679 selected feature combinations was modeled using the LDA function from the Mass R package (Venables & Ripley, 2010). The LDA works by finding the linear combinations of features that best separate the classes. It maximizes the ratio of the between-class variance to the within-class variance to reduce data dimensionality while keeping class differences. LDA models take on two main assumptions. The first one is that the data within each class has a Gaussian distribution. The second one is that the classes have the same covariance matrix. As shown in the previous chapter, these

assumptions should not be met for our data, given that not all feature distributions are normal. However, LDA is known for exhibiting good performance results even when these assumptions are unmet. As a matter of fact, these assumptions are seldom met in the wild. Figure 50 shows the results of the best models for each number of features:

Figure 50 - Overall accuracy and max accuracy score as a function of number of features



The first observation is that, at this stage, the accuracy score is measuring not the models' generalization capability but the goodness of fitness, i.e., how well the models fit the data. This means that each model was trained on the whole dataset and tested on the whole dataset. Models with more robust generalization power, using stratified k-fold and Leave-One-Out cross-validation sets, are evaluated further ahead.

Some observations are noteworthy. First of all, the overall

model's goodness of fit starts to stabilize (around 0.7 accuracy score) approximately from 13 n features on. This is in line with Mallow's Cp stat results, which indicated an optimal number of features around 15 (See Figure 48). However, the number of features needed for each class in the overall model vary a lot. The CNT class needs only two features. The INP class achieves a good fit with approximately three features. To achieve almost 0.88, the EXP needs only five features. The EVD class needs around 8 features to stabilize, and the ALL class needs around 7 features. The overall model needing 15, it is most likely that the features mobilized by each class are not always the same. This is checked in the following subsections.

Figure 51 below shows the LDA plot of the model using the 15 most selected features. An LDA plot is typically used to show the separation between classes in a four-dimensional space visually. Each axis corresponds to one of the four discriminant functions, which are linear combinations of the original features. The plot shows how well the classes are distinguished based on these linear combinations. It helps identify patterns and relationships in the data and showcase the effectiveness of LDA model in reducing dimensionality while preserving class-related information. Here, each class is shown with its label and a different color. We can see, for instance, that the least separated classes are overall EXP and INP. The most separable class, on the other hand, is ALL, a result that is in line with the results of the LDA model that took as input only the f_0 curve coefficients.

Figure 51 - LDA plot for 15 features

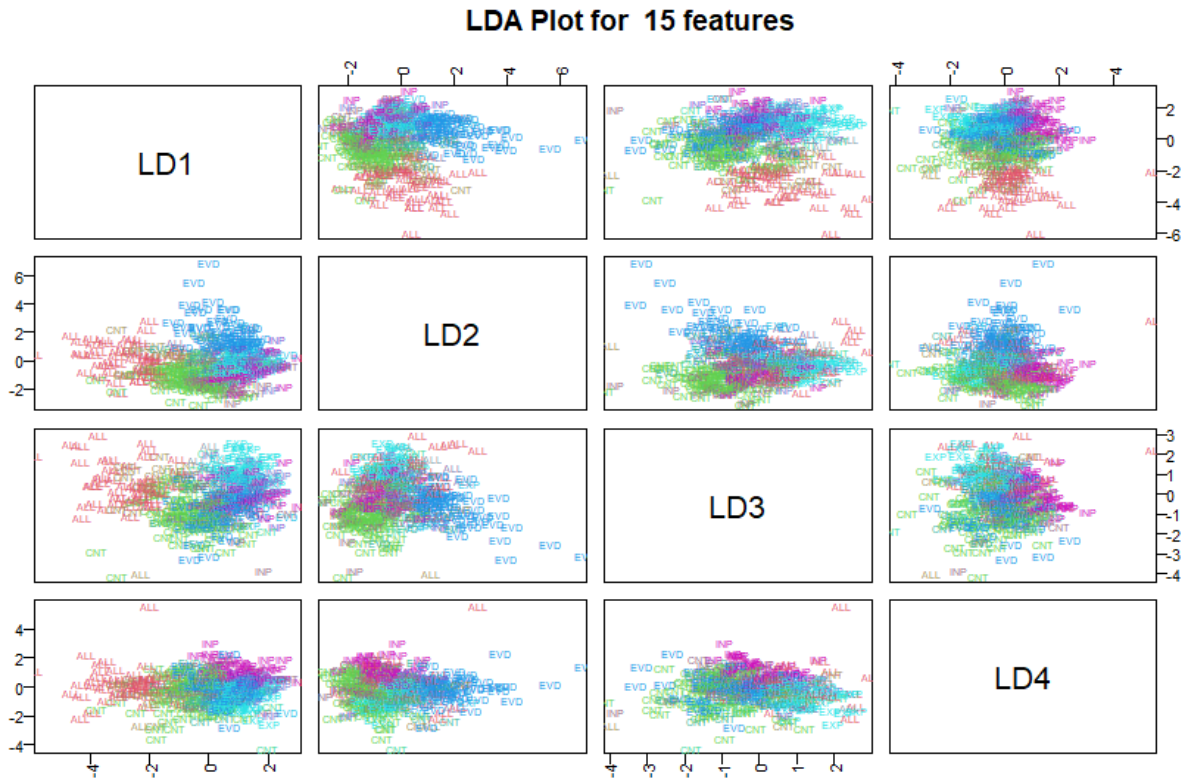


Table 33 displays the coefficients of the LDA. They stand for the weights assigned to input features projections onto the discriminant axes. The bigger the absolute value, the larger the effect on class separation. The sign of the coefficient (positive/negative) indicates the directionality of the feature's influence on class separation in the linear combination defined by the LDA. At the bottom of the table, the proportions of trace are also presented for each Linear Discriminant (LD). They refer to the eigenvalues of the covariance matrix, which represents the amount of variance explained by each LD.

Table 33 - Coefficients of discriminative functions

Feature	LD1	LD2	LD3	LD4
coef_1	2.642	1.003	1.768	0.720
coef_2	1.828	-0.074	0.370	-0.807
ratio_min_f0_stressed_dm	1.138	0.821	0.830	2.422

Feature	LD1	LD2	LD3	LD4
coef_0	1.084	0.912	1.259	1.391
ratio_max_f0_dm	0.775	0.628	0.457	-1.078
zsil_stressed_syl	0.665	-2.722	0.687	1.374
mean_intensity_stressed_dm	0.599	0.075	0.298	0.428
pitch_slope_stressed_dm	0.403	0.373	0.950	0.048
coef_3	0.371	-0.070	-0.484	-0.250
mean_pitch_dm	0.303	0.320	-0.404	-0.189
ratio_max_intensity_dm	0.256	-0.137	-0.157	0.079
ratio_min_intensity_dm	0.203	0.021	0.017	1.061
mean_intensity_dm	0.069	-0.248	-0.072	0.341
pitch_slope_after_stressed_dm	0.010	0.036	-0.248	0.109
ratio_max_f0_stressed_dm	-0.006	-1.970	-1.237	4.276
std_intensity_dm	-0.058	-0.650	0.144	-0.579
mean_se_stressed_dm	-0.067	-0.293	0.150	-0.179
std_pitch_dm	-0.108	-0.127	0.257	-1.027
max_pitch_dm	-0.183	-0.108	0.344	0.652
max_intensity_dm	-0.247	0.004	-0.171	0.104
min_intensity_dm	-0.249	-0.559	0.273	-0.887
min_pitch_dm	-0.293	-0.460	0.205	-0.635
ratio_max_intensity_stressed_dm	-0.362	0.124	0.757	-1.127
ratio_min_f0_dm	-0.390	-0.359	-0.481	-0.839
zsil_mean	-0.412	2.810	-0.986	-1.746
pitch_slope_dm	-0.468	-0.229	-0.752	0.153
ratio_min_intensity_stressed_dm	-0.581	0.255	0.156	-5.430
dm_duration	-0.667	-0.071	0.739	-0.095
PROPORTION OF TRACE	0.464	0.328	0.150	0.058

Together, LD1 and LD2 accounts for the 77% of the variance between classes. The 14 features with the most impact over each LD are indicated in red. For LD1, f0 curve coefficients, as well as relative duration and alignment of min intensity with respect to the stressed vowel play the most relevant roles. For LD2, mean syllabic duration and syllabic duration of the stressed syllable have the biggest impact. LD3 also has a non-negligeable impact on separating classes. The most

relevant features are curve coefficients and alignment of max f0 with respect to the stressed syllable.

Finally, Table 34 and Table 35 display, respectively, the confusion matrix and performance metrics by class for the 15-feature overall model. It is possible to see that the most separable classes are ALL and EVD, and the least EXP and INP.

Table 34 - Confusion matrix - Overall model

		Observation					Ratio				
		ALL	CNT	EVD	EXP	INP	ALL	CNT	EVD	EXP	INP
ALL	True label	53	14	0	0	0	0.791	0.209	0.000	0.000	0.000
CNT		8	101	9	5	21	0.056	0.701	0.063	0.035	0.146
EVD		4	4	58	6	3	0.053	0.053	0.773	0.080	0.040
EXP		3	8	5	47	15	0.038	0.103	0.064	0.603	0.192
INP		0	12	3	11	41	0.000	0.179	0.045	0.164	0.612
		Prediction									

Table 35 - Performance metrics by class for the overall model

METRIC	ALL	CNT	EVD	EXP	INP
Sensitivity	0.78	0.73	0.77	0.68	0.51
Specificity	0.96	0.85	0.95	0.91	0.93
Pos Pred Value	0.79	0.70	0.77	0.60	0.61
Neg Pred Value	0.96	0.87	0.95	0.94	0.89
Prevalence	0.16	0.32	0.17	0.16	0.19
Detection Rate	0.12	0.23	0.13	0.11	0.10
Detection Prevalence	0.16	0.33	0.17	0.18	0.16
Balanced Accuracy	0.87	0.79	0.86	0.80	0.72

7.4.2 ALL Against OTHERS

In this and the following subsections, I tested the same feature selection procedure with models trained to classify only two classes: a target class (here, ALL) and the label OTHER. To do that, all the observations from the target class were selected, and the other labels

were labeled as OTHER. This caused the classes to be highly imbalanced. I first tried to balance the data by selecting the same number of observations of OTHERS and the target DM class. This has proved better in terms of goodness of fit, but the resulting models were way too dependent on the observations randomly picked from the whole dataset. Setting different seeds might result in different features being selected as the most important. Moreover, the balanced subset would not reflect the decision-making process of the overall 5-class model. To remedy the imbalancedness, the best models were selected not by their global accuracy score but by the f1-score, which is a robust metric for imbalanced data. The goal was to find what features are more frequently used by the model when one class is checked against the others, as well as the optimal number of features. They are probably the best candidates to distinguish the target class from the others. This is useful, especially considering that the number of features necessary to achieve the best fit differs from class to class, as seen in Figure 49. In addition to the LDA models, a Decision Tree (DT) model (R rpart package - Therneau et al., 2015) was run using the same features selected by the LDA feature selection evaluation. The respective DT plot was also generated. A DT plot is a visual representation of the decision-making process of a model, making it useful for feature selection by highlighting the features that contribute significantly to the model's performance. The Cp statistic, the most selected features, and the DT plot are provided for each one-vs-others model.

The Cp statistic indicates the number of features that will achieve the best fit for the ALL-vs-OTHERS model, which should be around 13. This is shown in Figure 52. The features selected most frequently are, in their turn, shown in Figure 53.

Figure 52 - Cp statistic (ALL vs OTHERS)

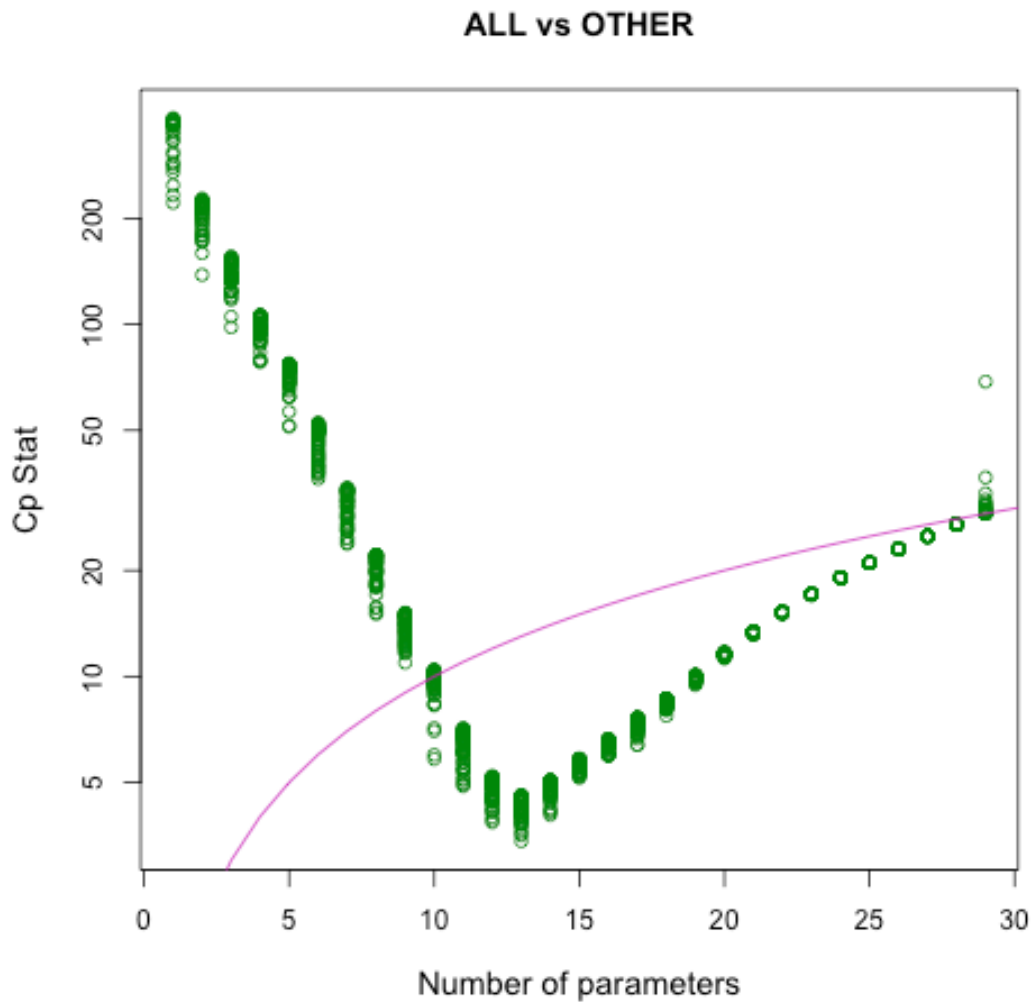
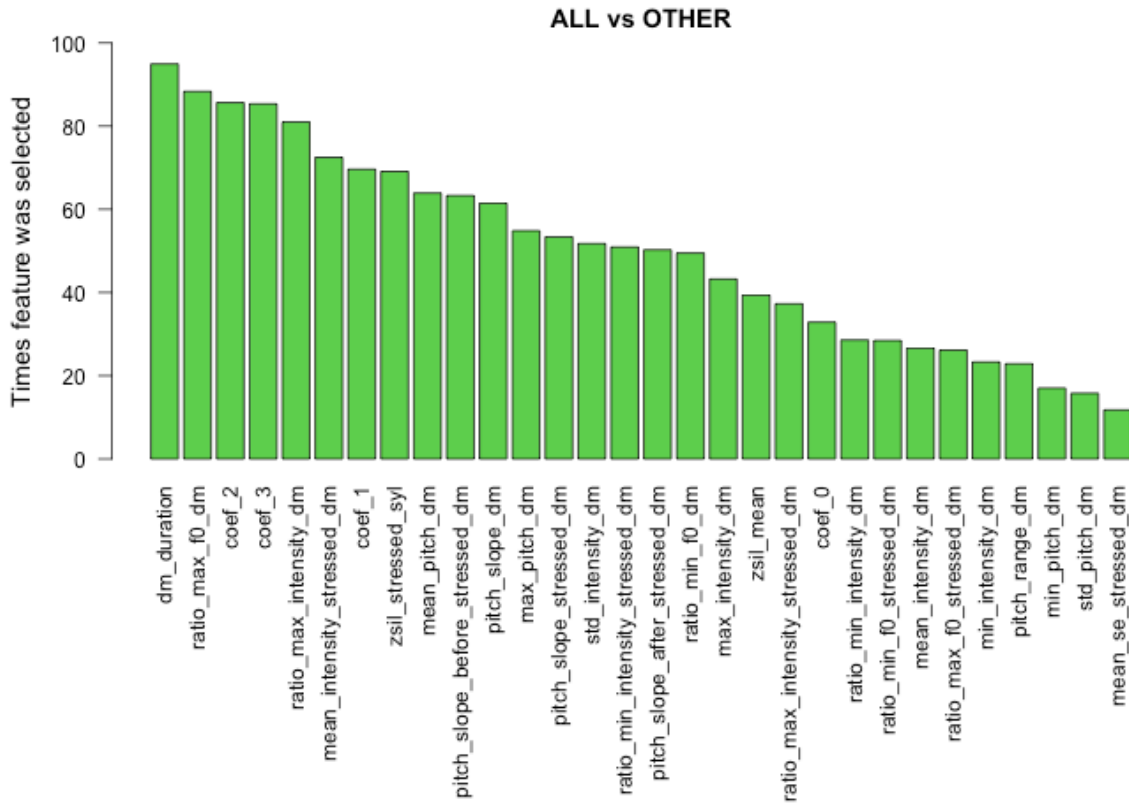


Figure 53 - Most selected features (ALL vs OTHERS)



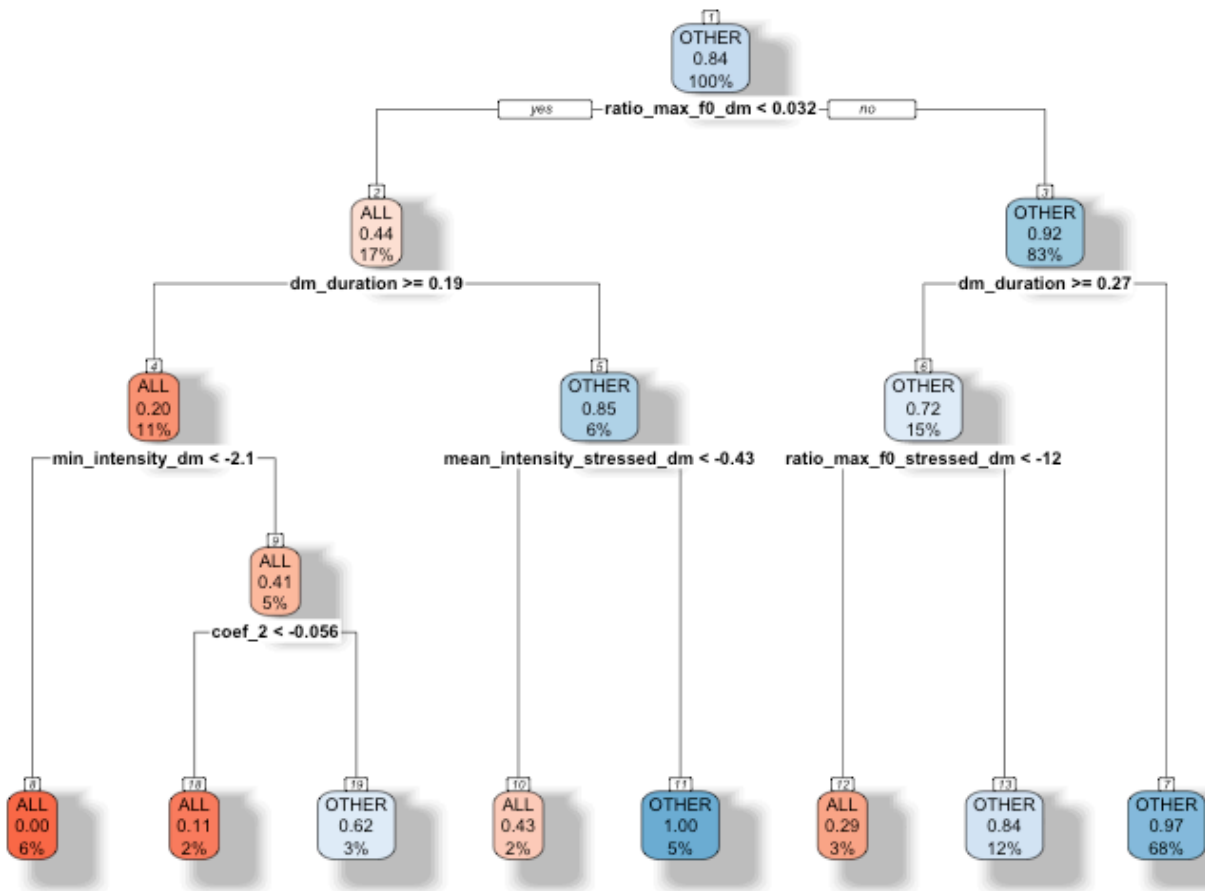
The model with the best f1-score fit only needs exactly 13 features. They are namely:

- a) Min intensity;
- b) Mean intensity in the stressed vowel;
- c) F0 slope in the whole DM instance;
- d) F0 range;
- e) F0 slope
- f) Alignment of max intensity;
- g) Alignment of max f0;

- h) Alignment of max f0 with respect to the stressed vowel;
- i) Mean syllabic duration;
- j) Relative duration;
- k) Second, third, and fourth f0 curve coefficients;

Notice that the best model will not necessarily have the optimal number of features indicated by the Cp statistic. It is also noteworthy that ALL selects almost all f0 curve coefficients. This was foreseeable since this is the DM class with the most distinctive curve, as seen in part 7.2. Curve fitting: ALL achieved the best classification scores based solely on the f0 curve. The model also selects intensity, duration, and f0 slope parameters as important predictors. ALL is the unit with the highest mean intensity; it is longer than EVD and CNT (which also occur in final positions), and it has a negative slope in pre-stressed syllables (when they exist). The DT plot in Figure 54 allows the visualization of the most important parts of the decision-making process.

Figure 54 - Decision Tree plot (ALL vs OTHERS)



Here, max f0 should be aligned as closely as possible to the initial boundary - f0 curve falling right from the start of the DM instance. DM duration is also important since ALL is longer than CNT and EVD. This is followed by the min intensity feature and the third f0 curve coefficient (coef_2). ALL exhibits the lowest intensity levels and the most distinctive f0 curve, as seen in Table 20. Table 36 displays the fit of the ALL-vs-OTHERS models.

Table 36 - Model fit (ALL vs OTHERS)

Metric	LDA	DTC
Accuracy	0.93	0.93
Avg accuracy	0.85	0.83
F1-score	0.78	0.75

7.4.3 CNT against OTHERS

Figure 55 and Figure 56 display the Cp statistic and the most selected features considering all evaluated CNT-vs-OTHERS models.

Figure 55 - Cp statistic (CNT vs OTHERS)

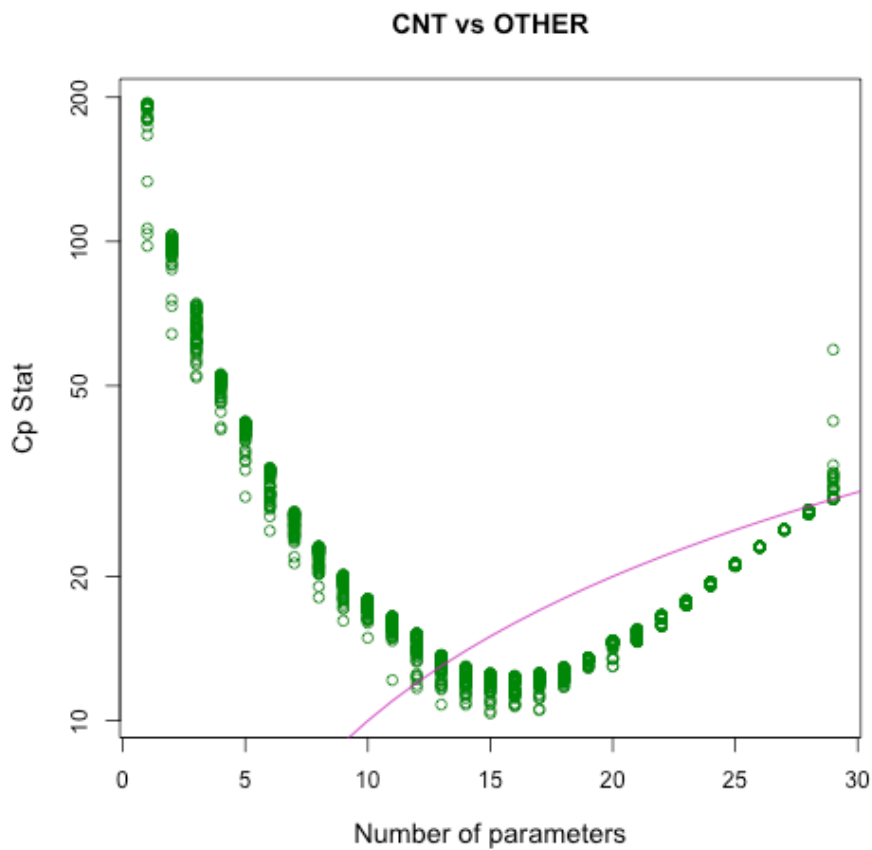
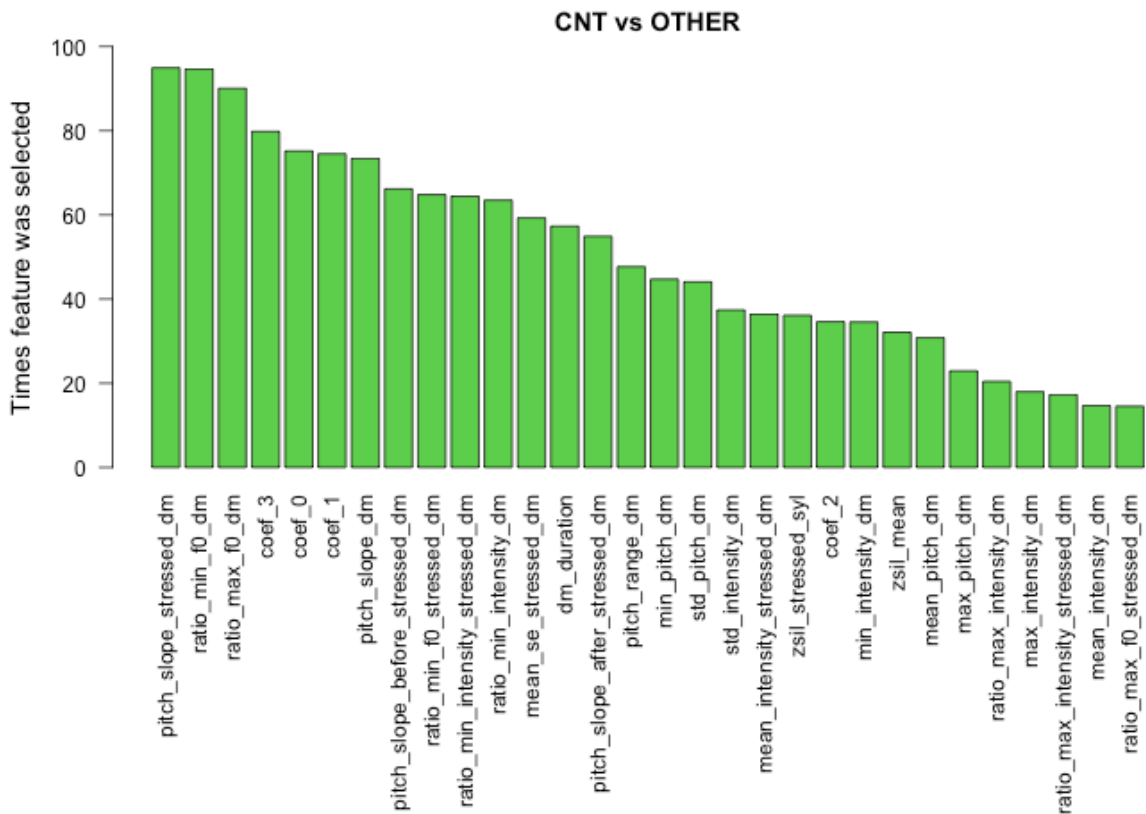


Figure 56 - Most selected features (CNT vs OTHERS)



The Cp statistic indicates that the optimal number of features should be around 15. The best accuracy is however obtained with 21 features. They are namely:

- a) Mean f0;
- b) Min f0;
- c) Intensity standard deviation;
- d) Min intensity;

- e) Mean intensity in the stressed vowel;
- f) Mean spectral emphasis;
- g) F0 slope on the DM instance;
- h) F0 slope in the stressed vowel;
- i) F0 slope before the stressed vowel;
- j) F0 slope after the stressed vowel;
- k) Alignment of max intensity;
- l) Alignment of min intensity;
- m) Alignment of min intensity with respect to the stressed vowel;
- n) Alignment of max f0;
- o) Alignment of min f0;
- p) Alignment of min f0 with respect to the stressed vowel;
- q) Mean syllabic duration;
- r) Relative duration;
- s) Third, second and fourth f0 curve coefficients;

The importance of these features can be observed in Figure 57. Here, a distinctive feature is the f0 slope in the stressed syllable. Indeed, CNT proved to have the highest f0 slope. While other units are characterized by rising and flat f0 movements, only ALL and CNT have sharply falling movements within the stressed vowel. The difference between these two units seems to be in the mean f0 level. As aforementioned, ALL has the lowest f0 level. This is not reflected in the DT plot, but the parameters of intensity alignment may play an important role in the distinction between these two units.

Figure 57 - Decision Tree plot (CNT vs OTHERS)

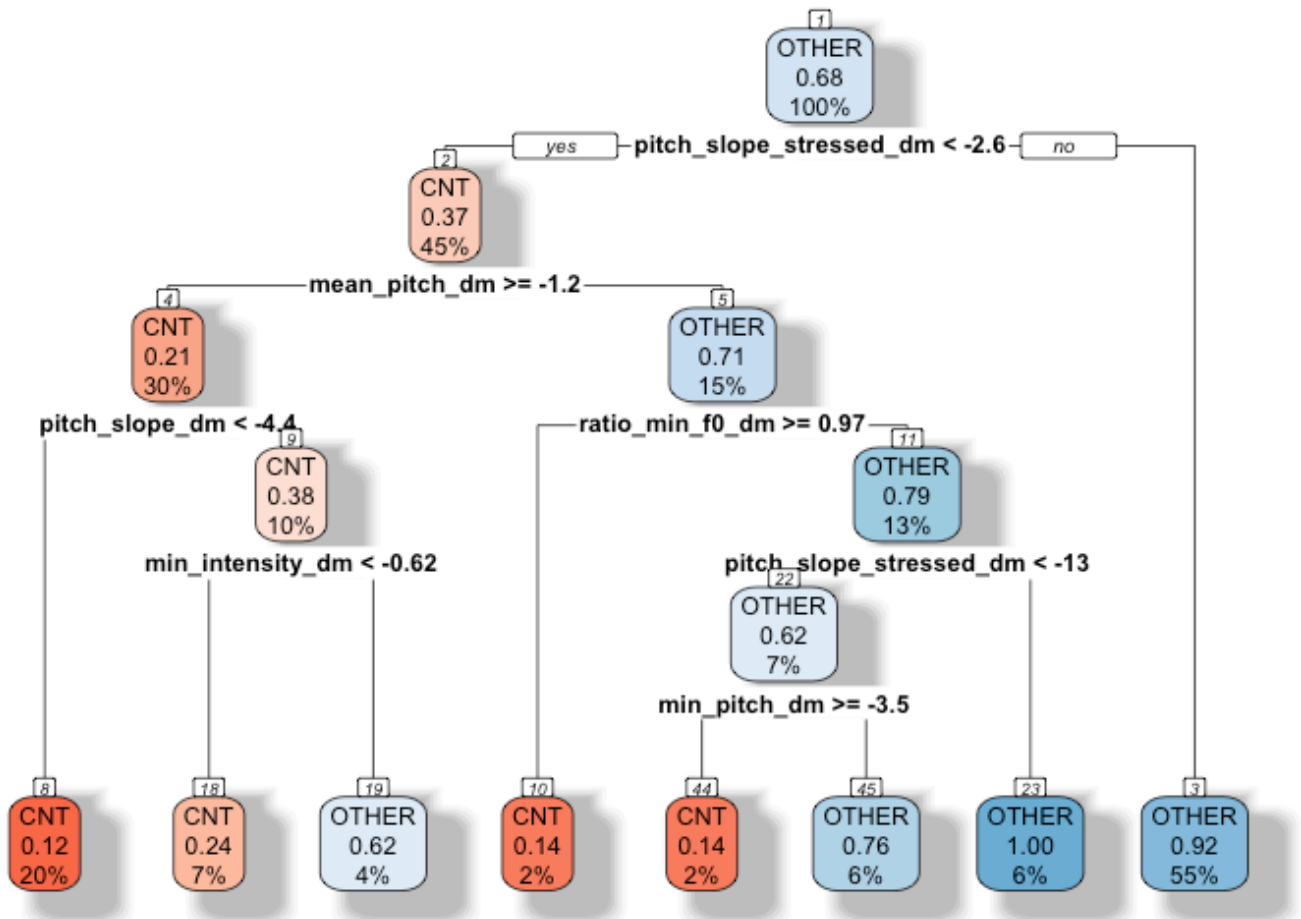


Table 37 below shows the fit of the best ALL-vs-OTHERS models.

Table 37 - Model fit (CNT vs OTHERS)

Metric	LDA	DTC
Accuracy	0.84	0.88
Avg accuracy	0.8	0.86

Metric	LDA	DTC
F1-score	0.73	0.82

7.4.4 EVD against OTHERS

Figure 58 and Figure 59 presents the Cp statistic and the most selected features by all EVD-vs-OTHERS models:

Figure 58 - Cp statistic (EVD vs OTHERS)

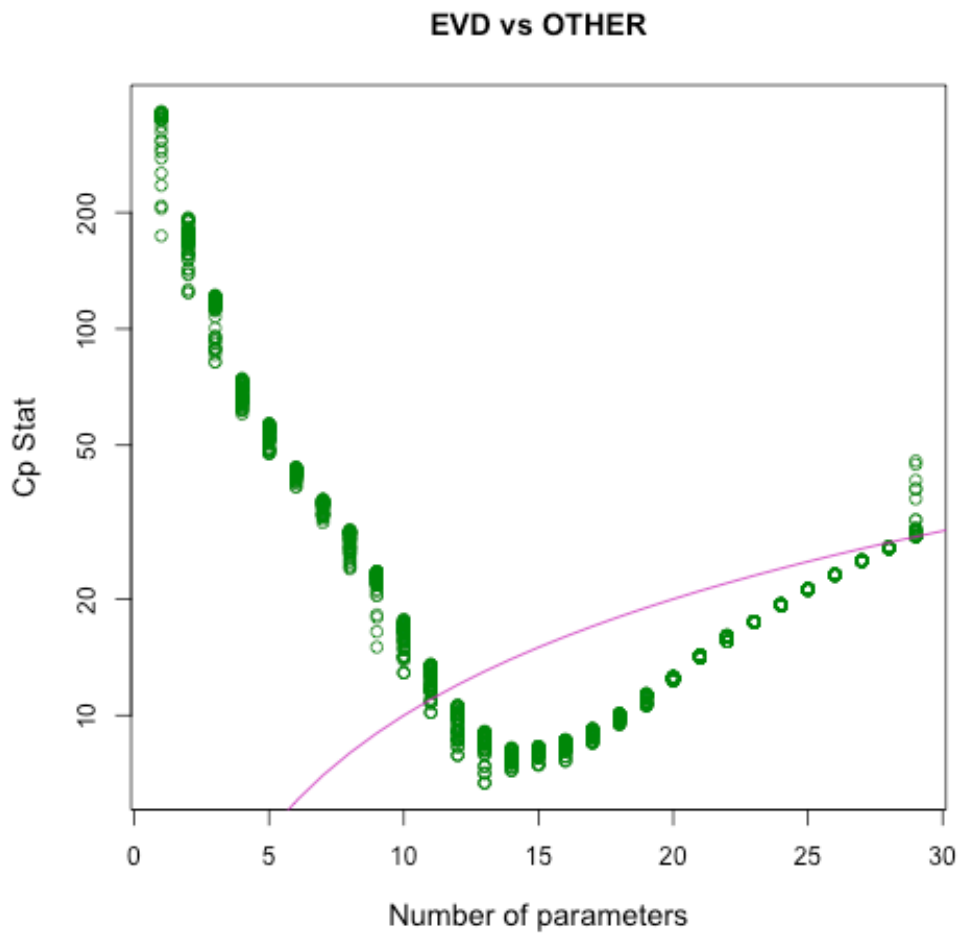
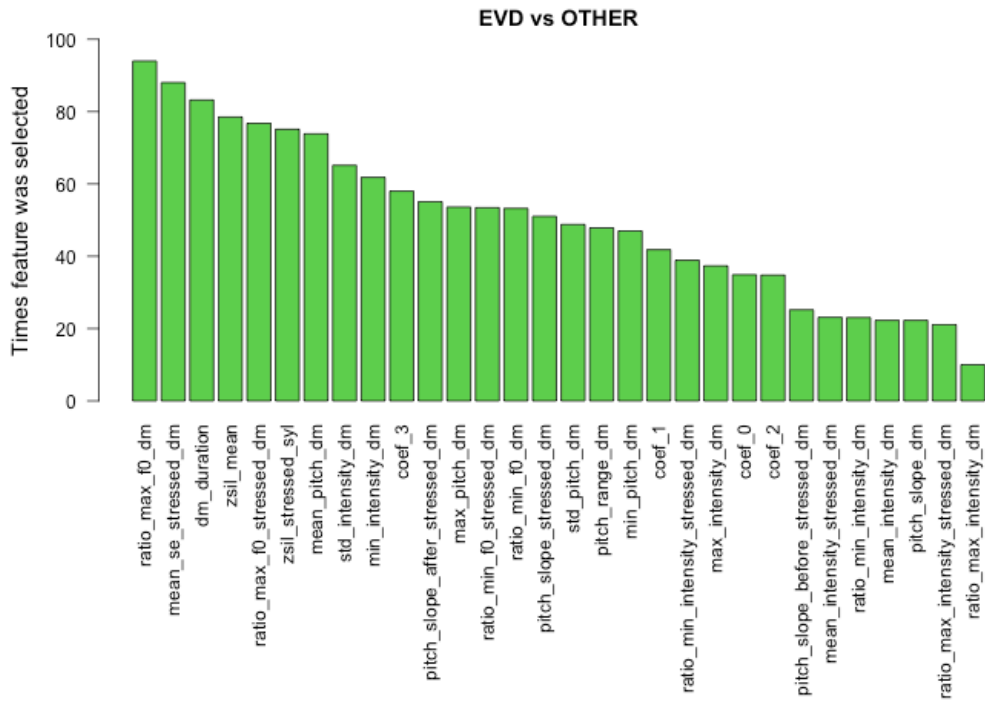


Figure 59 - Most selected features (EVD vs OTHERS)



The Cp statistic suggests that the best fit should be around 13 features. However, the best LDA model selects 21 features. They are namely:

- a) Mean f0;
- b) F0 standard deviation;
- c) Max f0;
- d) Min f0;
- e) Intensity standard deviation;
- f) Max intensity;
- g) Min intensity;
- h) Spectral emphasis in the stressed vowel;
- i) F0 slope in the stressed vowel;

- j) F0 slope before the stressed vowel;
- k) F0 slope after the stressed vowel;
- l) Alignment of min intensity with respect to the stressed vowel;
- m) Alignment of max f0;
- n) Alignment of min f0;
- o) Alignment of max f0 with respect to the stressed vowel;
- p) Alignment of min f0 with respect to the stressed vowel;
- q) Mean syllabic duration;
- r) Relative duration;
- s) Second, third and fourth f0 curve coefficients;

Against CNT and ALL, which also occur in the final position, EVD presents a rising f0 curve, with alignment of min and max f0 respectively at the beginning and at the end of the DM instance. Furthermore, these alignments tend to occur closer to the boundaries (opposite of CNT). The mean f0 level is also important to distinguish EVD from ALL. The former has a higher level, and the latter has the lowest f0 level. Figure 60 allows us to see that the parameters of alignment as well as the parameters of f0 slope in the stressed vowel play the most important role in the classification of EVD against other DM classes:

Figure 60 - Decision Tree plot (EVD vs OTHERS)

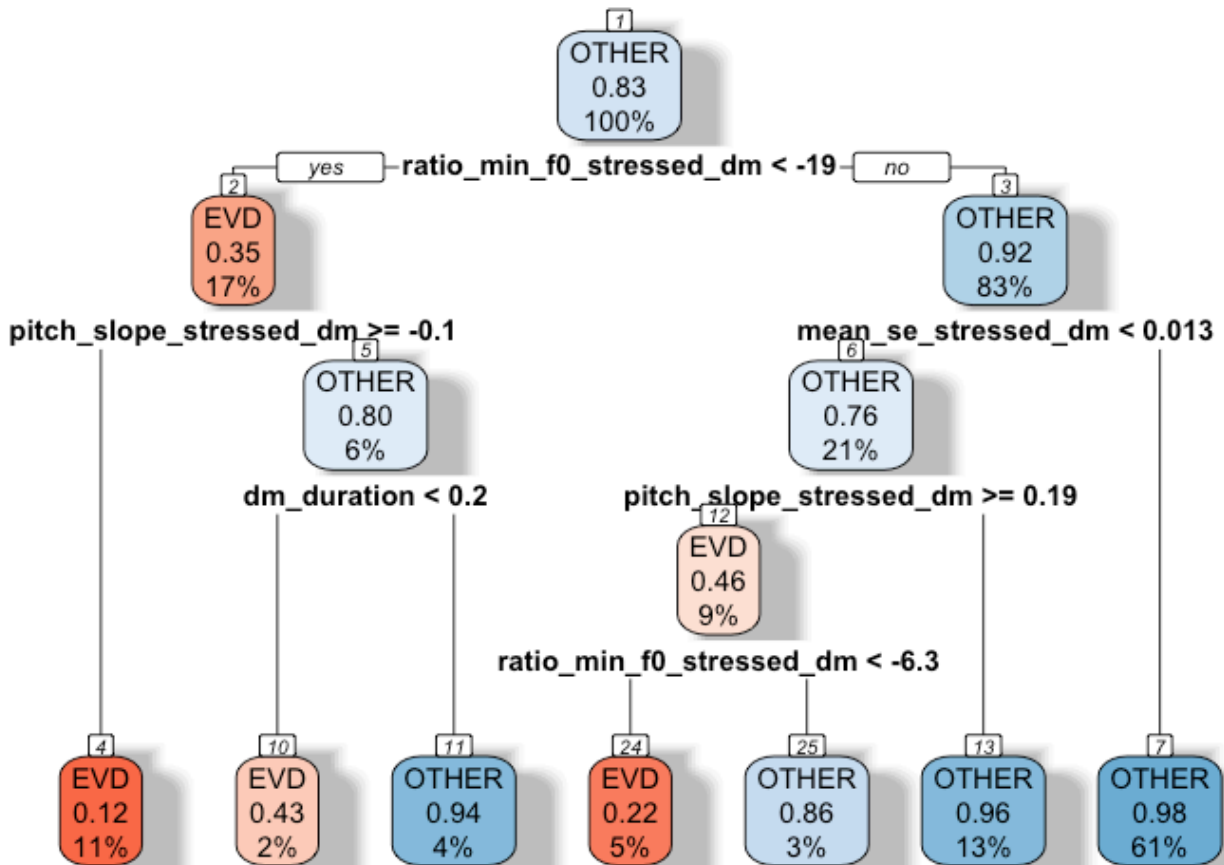


Table 38 exhibits the fits of the best EVD-vs-OTHERS models:

Table 38 - Model fit (EVD vs OTHERS)

Metric	LDA	DTC
Accuracy	0.93	0.94
Avg accuracy	0.85	0.91
F1-score	0.79	0.84

7.4.5 EXP against OTHERS

Figure 61 and Figure 62 show the Cp statistic and the most selected features for the EXP-vs-OTHERS models.

Figure 61 - Cp statistic (EXP vs OTHERS)

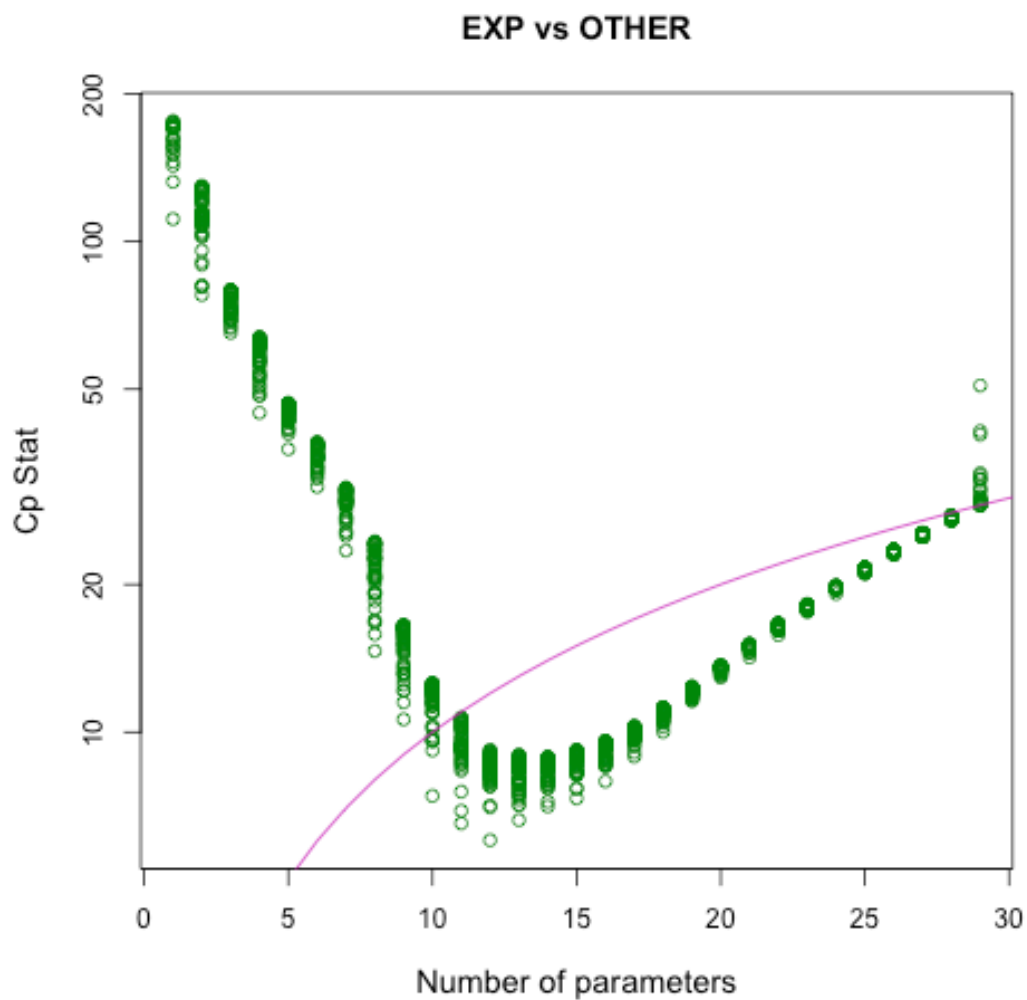
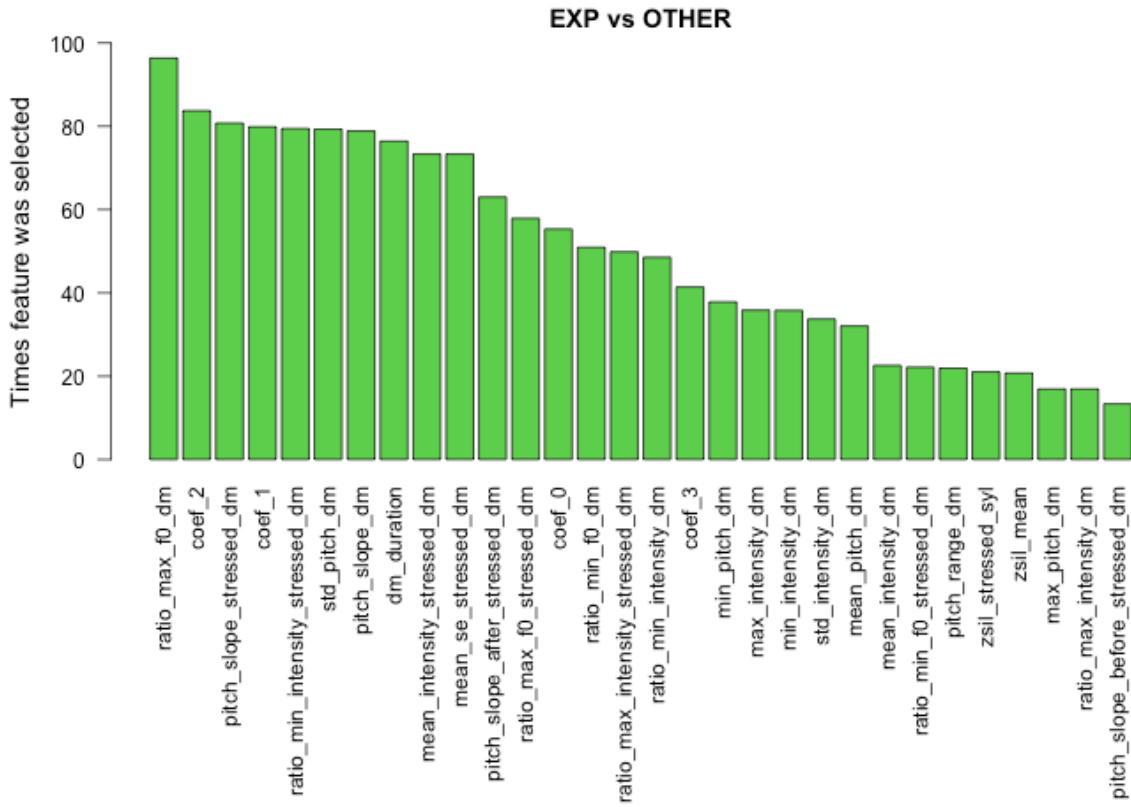


Figure 62 - Most selected features (EXP vs OTHERS)



The Cp statistic indicates that the optimal number should be met with 12 features. Again, the Cp value does not reach a value between 0 and 1. The best fit is also achieved with an exceptionally high number of features:

- a) F0 standard-deviation;
- b) Max f0;
- c) Intensity standard-deviation;
- d) Max intensity;
- e) Min intensity;
- f) Mean intensity in the stressed vowel;

- g) Spectral emphasis in the stressed vowel;
- h) F0 slope;
- i) F0 slope in the stressed vowel;
- j) F0 range;
- k) F0 slope after the stressed vowel;
- l) Alignment of min intensity;
- m) Alignment of max intensity with respect to the stressed vowel;
- n) Alignment of min intensity with respect to the stressed vowel;
- o) Alignment of max f0;
- p) Alignment of min f0;
- q) Alignment of max f0 with respect to the stressed vowel;
- r) Relative duration;
- s) First, second, third and fourth f0 curve coefficients.

The DT plot (Figure 63) helps explain what seem to be the most important features. Here, the first and most important one is the alignment of max f0 with respect to the stressed vowel. Values greater than 0 indicate that the max f0 point occurs after the central point of the stressed vowel, which is precisely what should happen with a unit that displays a rising f0 movement. The relative duration also plays an important role. As seen in the previous chapter, EXP tends to be the longest DM class. Furthermore, f0 slope in the stressed vowel reflects EXP's rising movement. This is also one of the most important features.

Figure 63 - Decision Tree plot (EXP vs OTHERS)

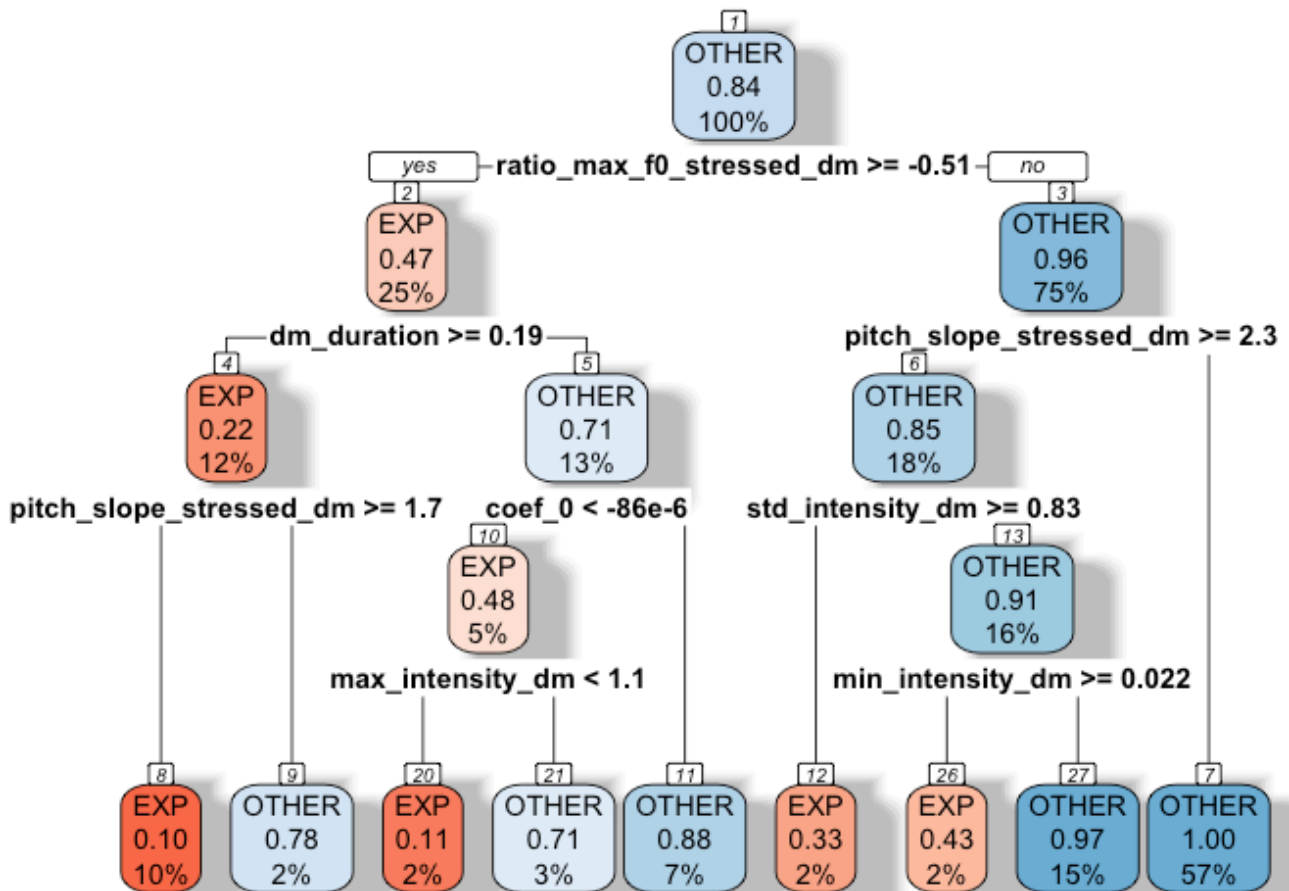


Table 39 shows the fit of the best EXP-vs-OTHERS models:

Table 39 - Model fit (EXP vs OTHERS)

Metric	LDA	DTC
Accuracy	0.9	0.94
Avg accuracy	0.77	0.89

Metric	LDA	DTC
F1-score	0.66	0.82

7.4.6 INP against OTHERS

Finally, Figure 64 and Figure 65 shows the Cp statistic and the most selected features of the INP-vs-OTHERS models. The Cp statistic reaches a value between 0 and 1 around 14 parameters:

Figure 64 - Cp statistic (INP vs OTHERS)

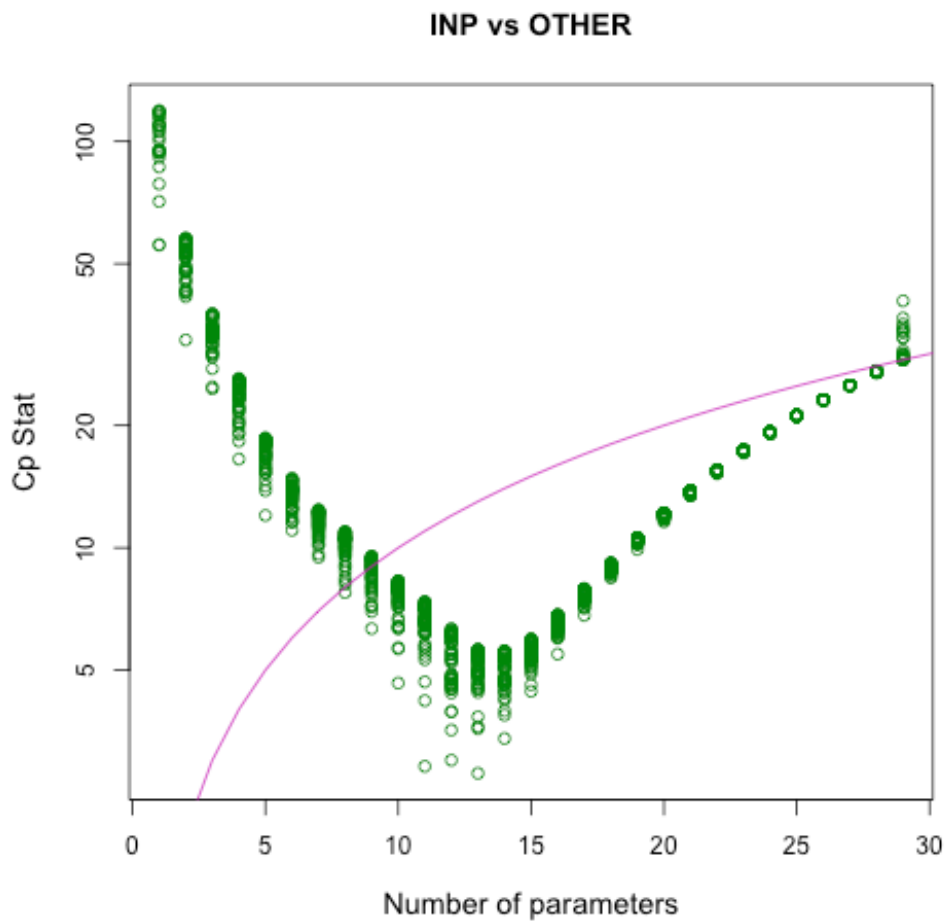
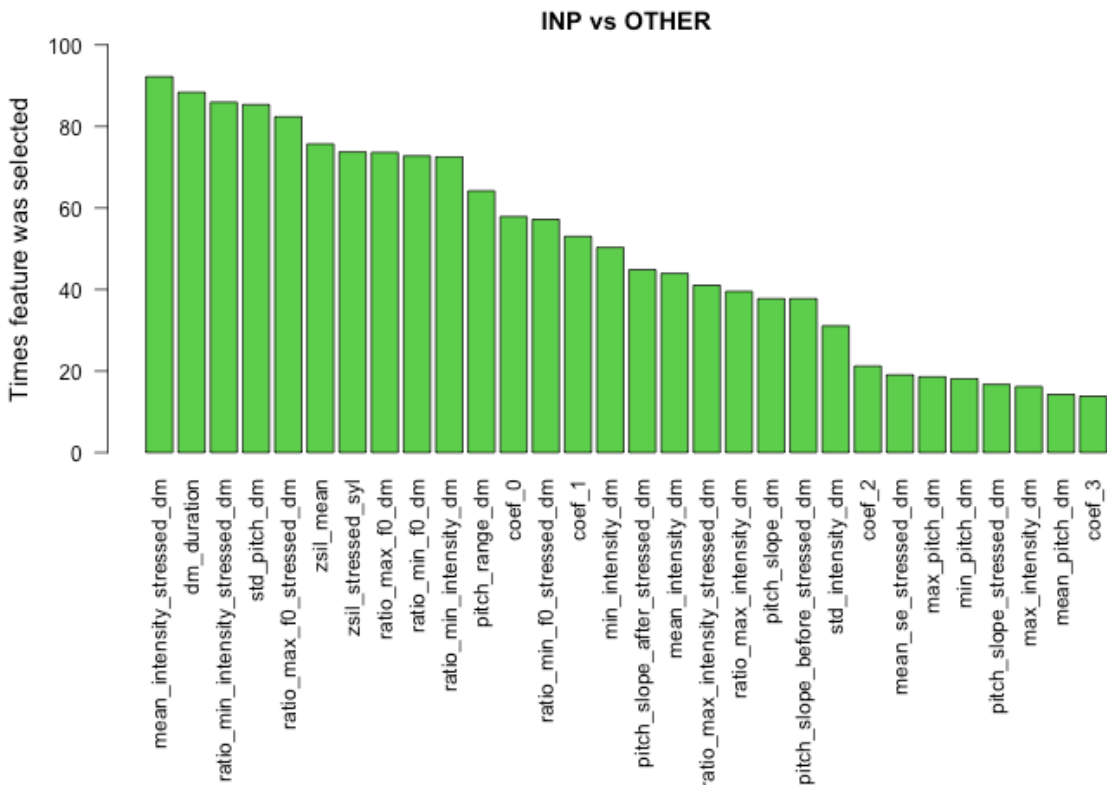


Figure 65 - Most selected features (INP vs OTHERS)



The LDA model with the best fit also selects many features (16). However, this model’s performance is equivalent to random guessing. Its f1-score is 0.5, as shown in Table 40. On the other hand, the DT model using the selected features displays an f1-score of 0.73. The most important features of this model are shown in Figure 66. Here, some combinations of features seem to be important. First, higher levels of mean intensity with a shorter f0 range in the stressed vowel combine with an f0 slope in the stressed vowel that should not surpass -1.5. Second, short relative duration combines with an f0 slope in the stressed vowel that should be greater than 3.7. The alignment of min f0 also seems to play an important role in the distinction. Interestingly, INP displays the most variability in terms of the f0 curve. It tends to display the flattest f0 movement in the stressed vowel, but it can be accompanied by rising and falling movements depending on the segments present before and after the stressed vowels. Since other DM classes in the same position can exhibit similar movements, the task of

classifying INP based on f0 shape can become more complicated.

Figure 66 - Decision Tree (INP vs OTHERS)

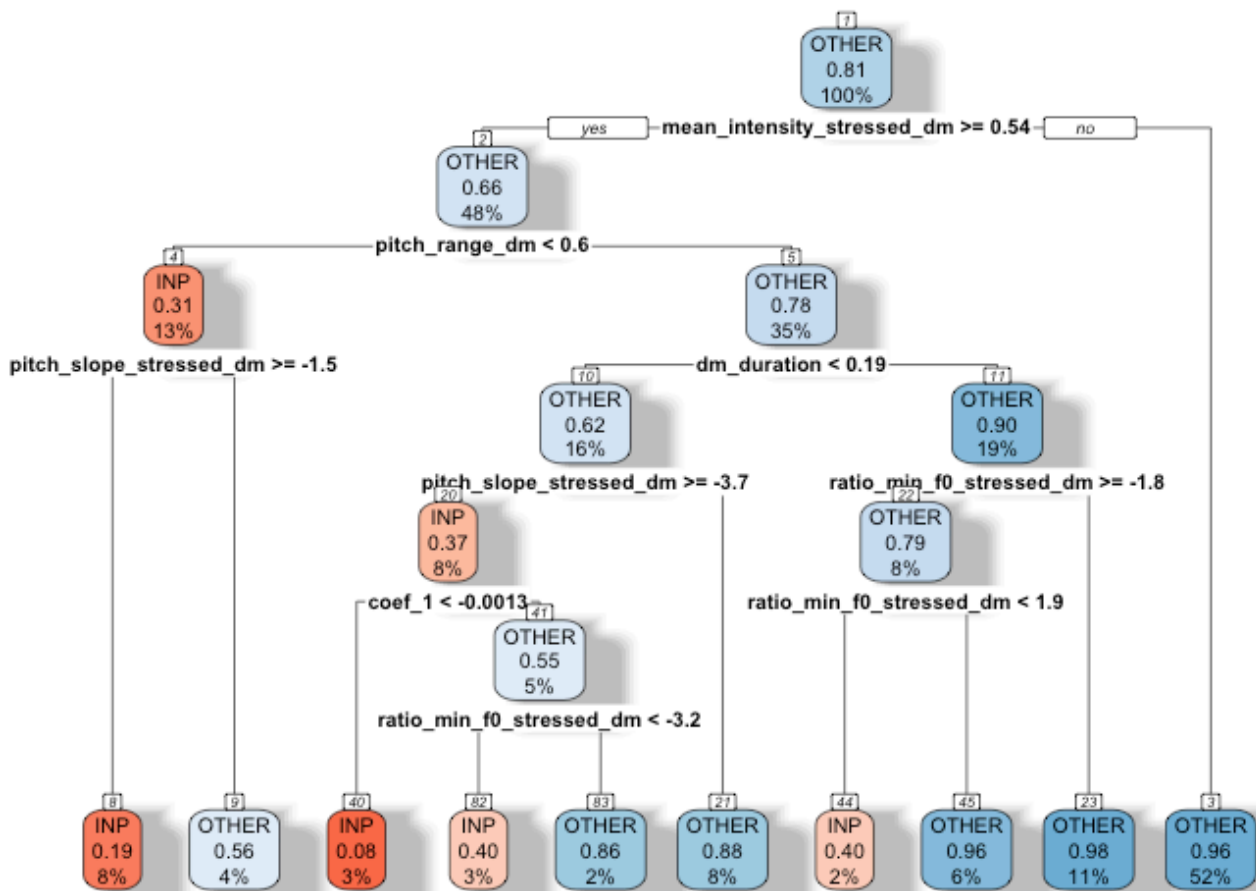


Table 40 shows the fit for the best INP-vs-OTHERS models.

Table 40 - Model fit (INP vs OTHERS)

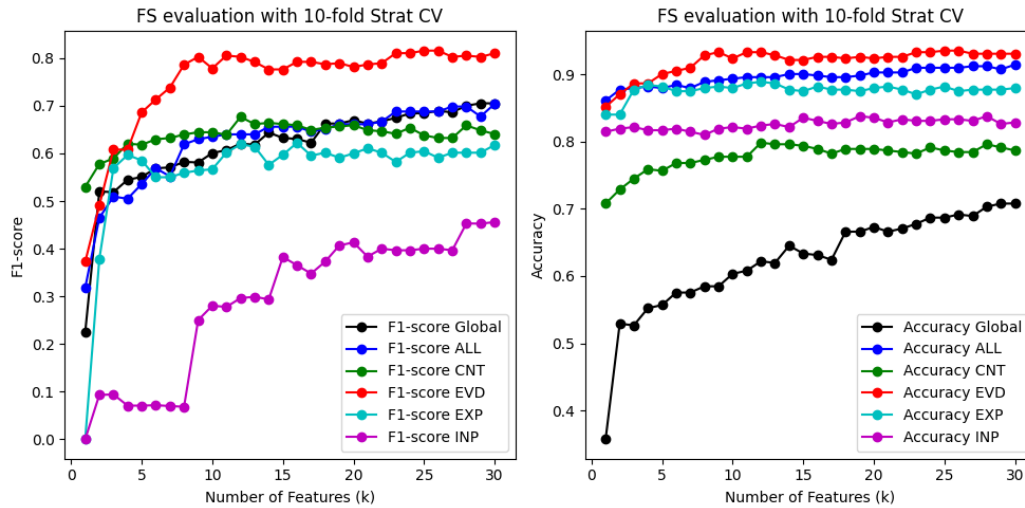
Metric	LDA	DTC
Accuracy	0.86	0.9
Avg accuracy	0.67	0.82
F1-score	0.5	0.73

7.4.7 Global model with Select K Best

Gobbo (2019) observed that as few as 9 features could be used to achieve a good model fit. The author's classification model took as input three DM classes, ALL, CNT, and INP, making up a total of 156 observations. In this dataset, INP always occurred in final position, CNT mostly in initial and final position, and ALL always in final position. The selected parameters were intensity, minimum intensity, alignment of min intensity, mean f0, alignment of maximum f0, f0 slope in the stressed syllable, number of syllables and raw duration. The accuracy score (goodness of fit) reached 84.6%. The best model seems to be a bit more intricate and convolute in our data. At some point, all features were selected by one of the models. I tried another feature selection approach using a different evaluation strategy to double-check the results shown in the previous subsections. The SelectKBest algorithm from Scikit-Learn (Pedregosa et al., 2011) removes all but the k features with the highest scores based on a specified statistical test or scoring function. Here, the F-statistic was used (f_classif method in Scikit-Learn). A stratified 10-fold cross-validation set was used to evaluate the best models. The average accuracy scores and f1-score were calculated from each fold. The results are shown in Figure 67:

Figure 67 - F1-score and accuracy score as a function of number of features resulting from the SelectKBest algorithm using a stratified 10-

fold cross-validation set



The results are quite consistent with the Leaps and Bounds approach. Increasing the number of features improves the model's performance. However, this improvement is much less noticeable from $k=4$ onwards. Two other observations seem to be confirmed. The most easily classifiable DM class is ALL. Assessed against the others, INP is less than chance.

7.5 OTHER MODELS

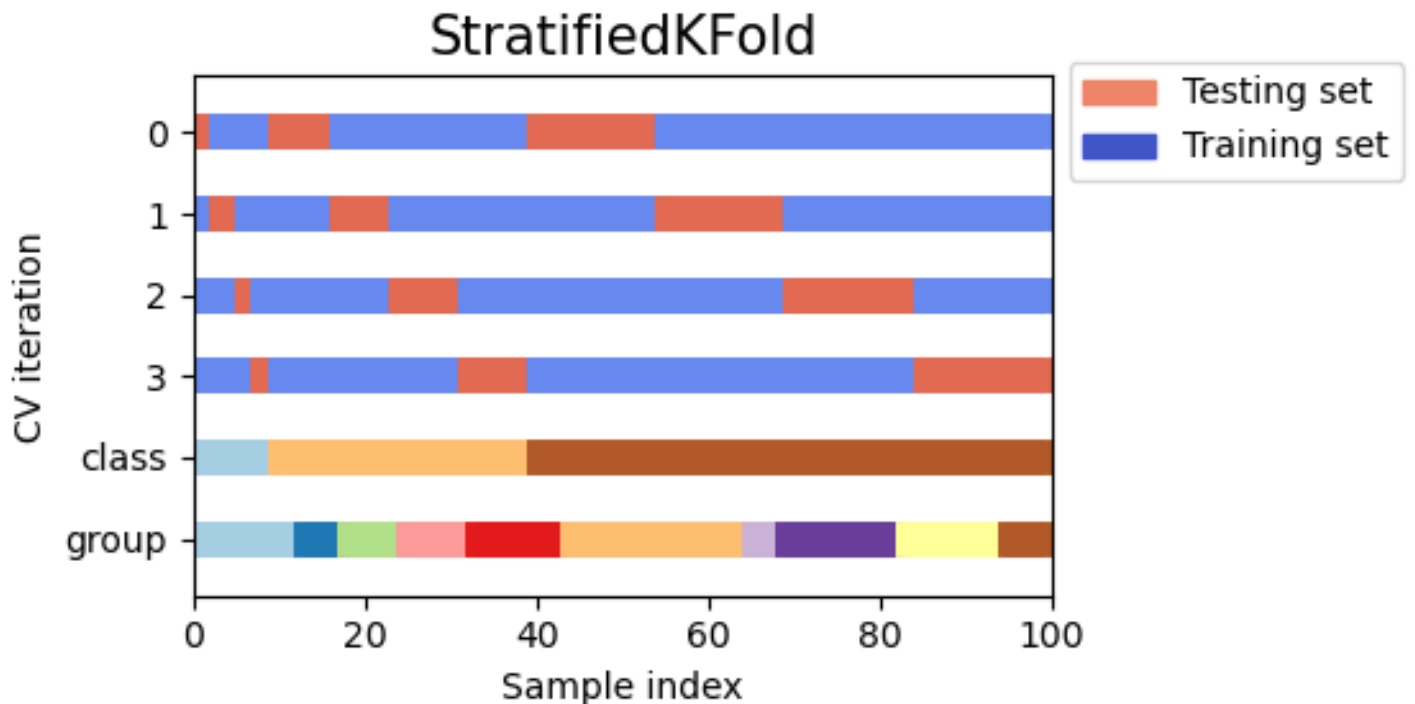
Three main modelling issues remain to this point. As mentioned before, the dataset used is imbalanced (1). This may cause some bias in results since the models presented so far may have a bias toward voting for the majority class (the CNT). Moreover, most of the models presented were not evaluated in a cross-validation set (2). This strategy has a number of advantages over classical evaluation strategies. It helps reduce bias by repeatedly training and evaluating the model on different subsets of the data. It can better predict how the model will perform in the wild since it will be repeatedly tested on unseen data. It helps detect and prevent overfitting, thus ensuring that the model is not learning unnecessary and way too many detailed patterns of the training data but is effectively learning useful patterns that generalize

well to new, unseen data. Finally, only non-finetuned, simpler classification techniques have been used thus far (3).

To address (1), two techniques aimed at rebalancing the data were employed: undersampling and oversampling. For undersampling, many techniques are available, the simplest being a random selection that takes the size of majority classes down to the size of the minority class. However, random undersampling can cause the loss of important information, especially when intra-class data presents high variance. For this reason, a NearMiss approach from the `imblearn` package (Lemaître et al., 2017) was preferred. NearMiss is also available in a large number of flavors. Here, the NearMiss-1 approach was chosen. NearMiss-1 focuses on reducing the number of observations of majority classes by picking samples that are close to the decision boundary of the minority class. It is thus focused on “hard-to-learn” observations, i.e., the observations of the majority class closer to the minority class. This is interesting because some DM classes, such as INP vs CNT, have hard-to-draw decision boundaries. For the oversampling, SMOTE (Synthetic Minority Over-sampling Technique) from the `imblearn` package was used. SMOTE generates synthetic observations of the minority class interpolating existing minority class observations.

To address (2), a stratified k -fold cross-validation approach (Pedregosa et al., 2011) was used. This cross-validation technique splits the dataset into k folds while keeping the same class distribution in each fold. The model is trained k times, with each fold as the test set once and the remaining folds as the training set. The final accuracy score is the average across all folds. This technique provides a more robust evaluation of the model's generalization capability.

Figure 68 - Visual representation of Stratified k -fold Cross-validation²⁹



Finally, to address (3), four classifiers (LDA, Decision Tree, Logistic Regression, and K-Nearest Neighbors) were fine-tuned. The best performing classifier was then used as a base estimator for a Bagging Model, which was further fine-tuned. The best bagging model was additionally evaluated on a Leave-One-Out Cross-Validation (LOOCV) set (from Scikit-learn, Pedregosa et al., 2011). LOOCV involves training a model k times, where k corresponds to the number of observations in the dataset. In each iteration, one observation is set apart to be used as the "test set", and the model is trained on the remaining observations. The final accuracy score is the average of the scores of these k evaluations. This technique provides a thorough

²⁹ Code available at < https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html#sphx-glr-auto-examples-model-selection-plot-cv-indices-py >

assessment of the model’s generalization capability with minimal bias. This technique was not used for the grid search on models’ hyperparameters because it would entail a very high computational cost.

The search spaces for the grid search of model’s best hyperparameters are not further described in this section, since they involve heuristics commonly used in Machine Learning approaches. These parameters, as well as the pipelines for this part of the work, are provided in a notebook. The results of the described methods are summarized in Table 41 below:

Table 41 - Results table for Bagging Models with balanced data

Sampling	Support	Best base estimator	BE's accuracy	Bagging model's accuracy	LOOCV accuracy
Undersampling	68	LDA	0.71	0.71	0.68
Oversampling	139	Logistic Regression	0.76	0.76	0.78

It is possible to say that the overall accuracy scores for this 5-DM-class classification task are somewhere between 68% and 78%, two issues considered. Firstly, that the data was undersampled with a more conservative approach, which may have caused some performance loss. Secondly, that oversampling with synthetic data is far from ideal. A general downside of the approach is that synthetic observations are created without taking into consideration observations from the majority class. This may potentially result in ambiguous observations when decision boundaries are fuzzy – strong overlap between classes (SMOTE, 2011).

Figure 69 and Figure 70, respectively for under- and oversampling, display the confusion matrices for the best bagging models considering the LOOCV approach:

Figure 69 - Confusion Matrix - LOOCV - Undersampling

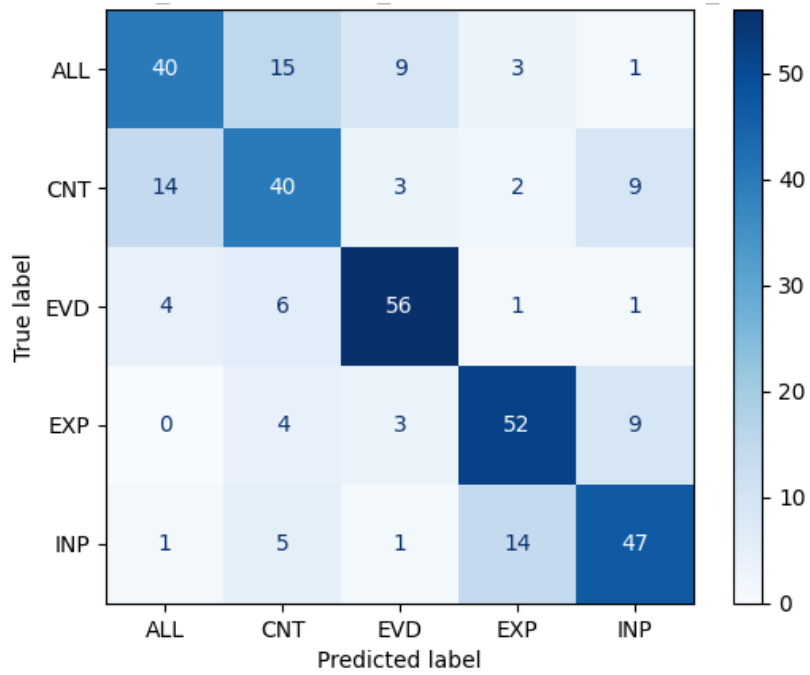
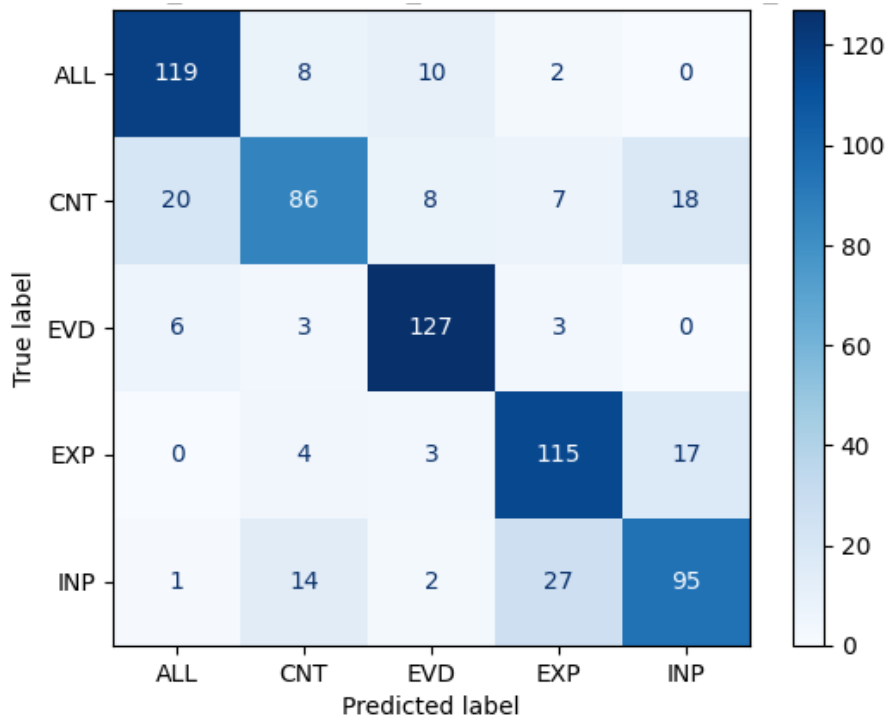


Figure 70 - Confusion Matrix - LOOCV - Oversampling



The most easily recognizable classes change from previous results. Before, INP was the most difficult and ALL the easiest DM class to classify. The EVD (Highlighter) is the most easily classified here, whereas CNT is often confounded with ALL and INP.

8 PERCEPTUAL EXPERIMENTS

8.1 INTRODUCTION

Chapter 7. (CLASSIFICATION MODELS) showed how and to what degree prosodic features can be used for distinguishing the functional nature of DMs within spontaneously produced speech utterances. This chapter presents a perceptual approach to evaluate the perceived differences across the functions of DMs in Brazilian Portuguese. It is based on the principles of the Language into Act theory (Cresti, 2000), which assumes that the prosodic form has a main role in implementing the DMs' pragmatic functions. Section 3.4. The most recent proposal for DMs presented five different DM classes, each implemented by prototypical prosodic contours. The prosodic implementation of these functions is also linked to other parameters, such as their position regarding the illocutionary unit (initial, medial, or final). Four functions (CNT, EXP, and INP) can occur at the beginning of the terminated sequence, but two of them, ALL and EVD, strongly prefer the final position in Brazilian Portuguese.

This chapter is a first step towards understanding how prosody can carry DM functions from a perceptive standpoint; the evaluation tasks presented in this chapter are restricted to the three functions that appear before the illocutionary unit: Incipit (INP), Conative (CNT), and Expressive (EXP). INP signals the speaker's intention to begin an utterance; EXP conveys non-illocutionary surprise; and CNT, which is distributionally free, indicates the illocutionary solution of the utterance. The goal is to evaluate the relevance of prosodic characteristics to implementing and perceiving these three functional categories. Another goal is to understand what other factors can contribute to or thwart the prosodic classification of DM functions. Two tasks are reported in this chapter. The first followed a discrimination paradigm with a restricted number of lexical fillers, and the second followed an identification paradigm with seven different lexical fillers.

The following sections present the selection of DMs, their

prosodic modification by a speech resynthesis procedure, and the implementation of the perceptual evaluations, using the paradigms mentioned, to evaluate the ability of participants to match the prototypical prosodic forms to a definition of each of these three functions.

8.2 DATASET OF THE DISCRIMINATION TASK

Three utterances were selected from the C-ORAL-BRASIL corpus (Raso & Mello, 2012) to present an illocution immediately preceded by a DM, the latter conveying one of the three targeted functions (CNT, EXP, INP). The original versions of these utterances were used for resynthesis and presentation to participants. This choice (keeping the original audio content) led to more restrictive data selection criteria. Among the criteria used to select the utterances were:

- a) The quality of the audio: the C-ORAL corpus contains spontaneous data that may have adverse recording conditions; this could impair the quality of prosodic modifications; thus, these stimuli were rejected.
- b) A lexical unit used as DM may have several functions, but not necessarily all the three targeted here, so only lexemes compatible with the three functions were selected.
- c) Some lexemes may also generate functional confusion to the listeners due to heavier semantic load; for this reason, the potential lexical items had to be restricted to lexemes as light as possible from a semantic standpoint.
- d) For similar reasons, the illocution following the selected DM must have a value adequate to the three functions.

The difficulty in finding examples fulfilling all these conditions explains the restricted set of examples used here; other instances were found, but their quality, as well as a preference to keep the experimental task as short as possible (here about 15 minutes), led us to keep only three utterances. These three utterances (produced by three different adult speakers, one male, for the INP, and two females, all speakers from the Minas Gerais variety of Brazilian Portuguese) were the following (with the DM enclosed in squared brackets):

- a) CNT: “[ah], não acaba não” (“[ah], it’s not over”)
- b) EXP: “[ah], primeiro a letra” (“[ah], first the letter”)
- c) INP: “[gente], é so um professor falando” (“[guys], it’s just a professor talking”)

8.3 DATASET OF THE IDENTIFICATION TASK

In the identification task, we also wanted to evaluate the effect of the lexical content of the DM. Thus, a more varied number of lexical fillers were selected from the C-ORAL-BRASIL corpus. Here, the quality of the audio was not an issue, because the examples were reproduced by one native speaker of BP in a controlled setting. The main restriction for this selection was that each lexeme should occur in all targeted functions within the corpus. Furthermore, preference was given to utterances that carried more neutral, assertive illocutionary units, so to avoid this uncontrolled factor having an effect linked with the specific speech act performed. The seven different lexemes chosen were: *ah* (oh), *é* (yeah), *gente* (guys), *não* (no), *oh* (oh/look), *porra* (fuck), *uai* (typically regarded as a mark of surprise/disbelief). Three examples per lexeme were selected, each one carrying a target function. The exceptions were *ah* and *uai*. For the former, six examples were chosen (two per DM function). This is because a total of 8x3 utterances were

required to complete the Latin square experimental design (see latter). However, only seven different lexemes were founded fulfilling the three targeted functions. For *uai*, one additional CNT was selected to be used as a training stimulus. Each example was manipulated and resynthesized to resemble as much as possible the prototypical forms of the target functions. The original examples were:

Table 42 - Examples used in the identification task

LEXEME	FILE	ORIGINAL FUNCTION	TEXT
AH	bfamdl04_132	CNT	ah / mas é claro //
			Oh / but that's obvious //
AH	bpubdl01_119	CNT	ah / não acaba não //
			oh / it doesn't end //
AH	bfamcv04_263	EXP	ah / primeiro a letra //
			oh / first the letter //
AH	bfamdl05_267	EXP	ah / ele vai colocar corrimão //
			oh / he's going to install handrails //
AH	bfamdl01_241	INP	ah / vão levar esse mesmo //
			oh / let's take this one //
AH	bfamdl01_260	INP	ah / mas esse é ruim //
			oh / but that's a bad one //
É/EH	bpubdl02_215	CNT	é / eu trouxe o oito e o nove //
			yeah / I brought the eight and the nine //
É/EH	bpubdl02_054	EXP	é / mas é mesmo //
			yeah / but that's right //
É/EH	bfamdl01_096	INP	é / hoje cê tá faminta //
			yeah / you're starving today //
GENTE	bfamdl26_67	CNT	gente / é muito bonitinho //
			guys / so cute //
GENTE	bfamdl03_35	EXP	gente / eu te falei //
			gosh / I told you //
GENTE	bfamcv26_262	INP	gente / é só um professor falando //
			guys / it's just a professor talking //
NÃO	bpubmn01_093	CNT	não / a diretora muito boa //

LEXEME	FILE	ORIGINAL FUNCTION	TEXT
			no / the principal is a very nice person //
NÃO	bfamcv05_87	EXP	não / vai sô //
			no / go man //
NÃO	bfamcv02_141	INP	não / mas ea tá é brincando //
			no / but she's just kidding //
OH/O'	bfamdl22_090	CNT	o' / tem um dinheiro preso aqui no banco //
			look / you have a balance stuck here at the bank //
OH/O'	bpubdl02_145	EXP	oh / o bondade sua //
			oh / how kind of you //
OH/O'	bpubdl11_292	INP	o' / ajudar ele ajuda //
			look / he does help //
PORRA/PÔ	bfamcv32_212	CNT	pô / só três minutos //
			fuck / just three minutes //
PORRA/PÔ	bfamdl20_188	EXP	pô / o cara tá famoso //
			fuck / the guy got famous //
PORRA/PÔ	bfamdl17_132	INP	pô / garçom underground //
			fuck / underground waiter //
UAI/UÉ	bfamdl21_047	CNT	uai / tem que ter isso aqui também //
			oh / there has to be that here too //
UAI/UÉ	bfamdl28_078	CNT	uai / vamo ver //
			well / we'll see //
UAI/UÉ	bfamdl33_105	EXP	uai / cê já pôs o trem pra fritar //
			oh / you're already frying this thing //
UAI/UÉ	bfamcv11_041	INP	uai / ele conversa demais da conta //
			oh / he just talks too much //

8.4 RESYNTHESIS

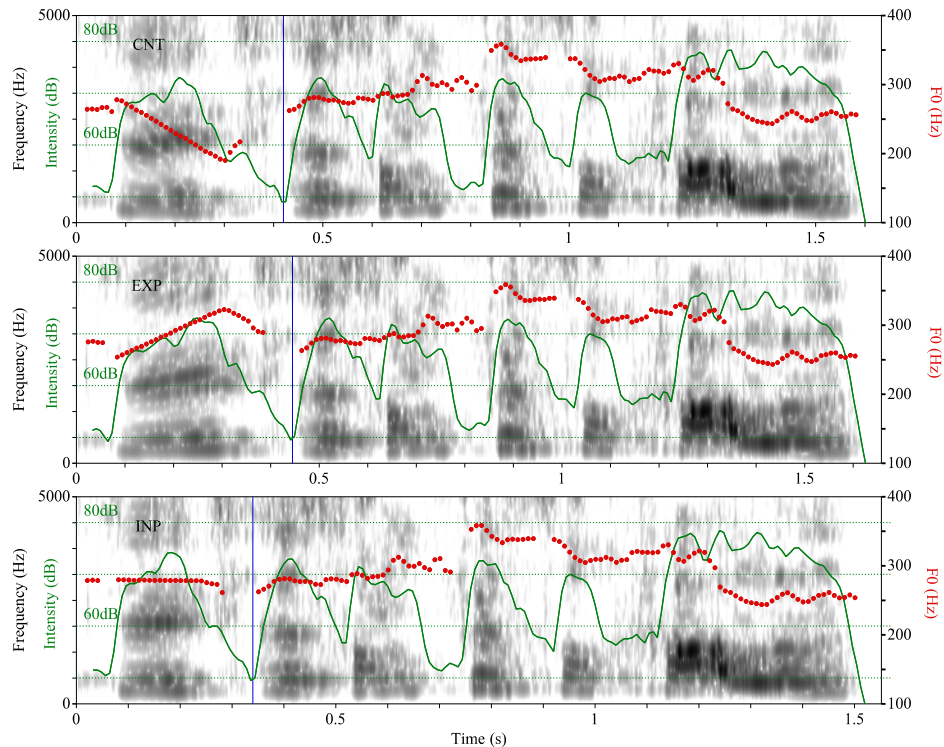
The utterances were extracted from the original recordings and edited in the following way. For the discrimination task, a noise reduction algorithm available in Praat (Boersma & Weenink, 2022) was applied to

the signal to remove some stationary background noises (cars passing in the street, notably). The illocutionary part of the utterance was then edited to keep only the targeted part; in that case, the final part of the illocution was also modified to sound like a terminated assertion (final pitch and intensity fall). This was not done for the audio files used in the identification task, since their reproductions were recorded in a controlled environment and the illocutionary units were realized with terminal boundaries.

For both tasks, the prosodic characteristics of the DM units were then modified to correspond to the prototypical description of the three targeted functions. These modifications were done by using Praat's "Manipulate" function, that allows varying speech fundamental frequency (F0) and duration using the TD-PSOLA algorithm (Moulines & Charpentier, 1990), and then by modifying the sound intensity, using Praat *IntensityTier* objects. Figure 71 shows the spectrograms, with overlaid f0 (red dots, in Hz) and intensity (green line, in dB) contours resulting of the modification process for the three prosodic functions (CNT, EXP, INP, from top to bottom) on the DM "gente" and the utterance "é so um professor falando". The end of the DM is marked with a blue vertical line.

Figure 71 - Example of manipulations of the three utterances used in the discrimination task – Red points represent f0 tracking and the

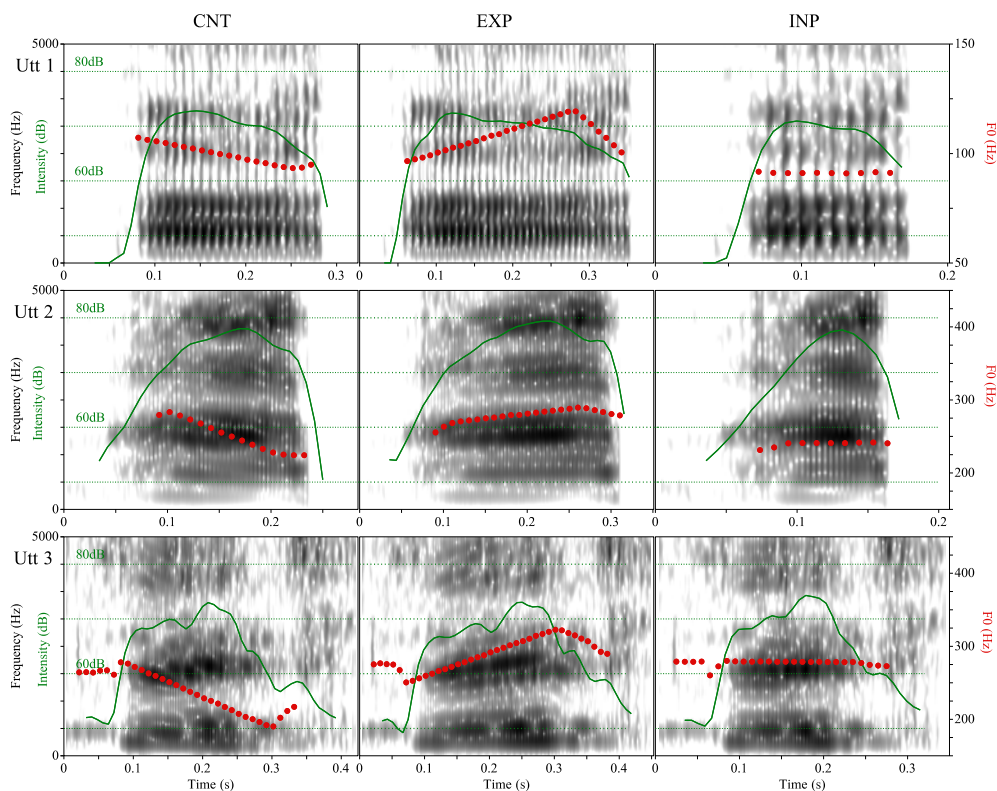
green curve represents intensity levels



The targeted prosodic characteristics linked to each function may be described in the following way (see sections 3.4. The most recent proposal for DMs and 7.2. Curve fitting). CNT displays a falling f_0 movement from the stressed vowel after a slightly rising movement, if there is pre-stress material; it has a much lower intensity and a shorter segmental duration than the illocution's mean. EXP presents a rising f_0 movement until the end of the stressed vowel, then a flat or slightly falling f_0 movement, if there is post-stress material; its segmental duration and intensity levels are below that of the illocution mean but above that of CNT. The INP has a flat f_0 , with a very short duration and higher intensity than the illocution. The original and modified versions of the DMs are presented in Figure 71 for the complete utterance "gente / é só um professor falando" (the visible desynchronization of the illocutionary part comes from the different durations of the three version of the "gente" DM). The DMs resulting from this process of

prosodic stylization to elicit the prototypical prosodic characteristics of the CNT, EXP and INP functions are presented in (without the associated illocutions), so as to make the prosodic similarities within a function more visible.

Figure 72 - Manipulated DMs of the discrimination task



The original DMs were each modified in three versions, leading to the stimuli that have a given lexical and segmental origin (the original DM in its illocutionary context), and three prosodic variants, that correspond to the three functions (CNT, EXP, INP). The exact prosodic characteristics of a DM for a given function may obviously vary with the speaker and the illocution characteristics (low or high vocal effort, for example, but the prototypical prosodic characteristics are those described above. This process resulted in nine stimuli for the discrimination task and 78 stimuli for the identification task.

Finally, for the identification task, the stimuli were also presented in their written form without any audio. The goal was to evaluate potential biases carried by the lexemes (without the prosodic realization).

8.5 PARTICIPANTS

Participants were recruited via social media to pass the perceptual experiments. All were adults, L1 speakers of BP. The tests were anonymous, and participants were asked to have the tests via a web interface. They were instructed to wear headphones, and an informed consent form was displayed, which they had to accept before starting. They should first answer three demographic questions (age, gender, and if their first language was BP) before initiating the experiments. Some participants connected to the interface but stopped the process before actually having the test (just answering the demographic questions); they were not included in the analyses. Table 43 and Table 44 show the summary of participants for the discrimination and the identification test, respectively:

Table 43 - Summary of participants of the discrimination test

Discrimination test			
	Total participants	Mean age	Std age
Female	53	28	13
Male	32	33	12.1
Total/Avg	85	29.9	12.8

Table 44 - Summary of participants of the identification test

Identification test			
	Total participants	Mean age	Std age
Female	68	27.7	13.6
Male	52	27.8	9.9
Total/Avg	120	27.8	12.1

8.6 DISCRIMINATION AND IDENTIFICATION PARADIGMS

The interfaces for the perceptual experiments were developed using the web-based "PsyToolkit" application (Stoet, 2010, 2017). It has been shown that online psychological evaluations do reach results similar to those of in-lab testing for a number of classical psychological evaluations (Kochari, 2019; Sasaki & Yamada, 2019). Therefore, the web-based interface was here preferred as it simplifies greatly recruiting participants with more varied profiles.

8.6.1 Discrimination task

The stimuli were presented in pairs: for the three versions of a DM from the same utterance, three pairs were made (CNT-EXP, CNT-INP, EXP-INP). For the three utterances, this leads to nine pairs of stimuli. For each pair presentation, participants had to judge which of the two prosodic contours of the DM best fit a given definition, in an AB discrimination protocol. The definitions that correspond to the three tested functions were the following (where "DM" was replaced by either "GENTE" or "AH", according to the tested lexeme):

CNT: *"Se você quisesse passar uma ideia de conclusão em função do que foi dito antes, qual das duas realizações de DM você escolheria?"* ("To convey an idea of conclusion based on what was said before, which of the two performances of DM would you choose?")

EXP: *"Se você quisesse manifestar que ficou surpreso com o que foi dito antes, qual das duas realizações de DM você escolheria?"* ("To express that you were surprised by what was said before, which of the two performances of DM would you choose?")

INP: "*Se você quisesse apenas começar a frase, qual das duas realizações de DM você escolheria?*" ("If you just wanted to start the sentence, which of the two performances of *DM* would you choose?")

Each pair was presented twice, with alternatively one of the two definitions corresponding in turn to one of the two prosodic versions of the pair, leading to 18 presentations (of pairs plus definition) for each judge. For each of the 18 presentations, the pairs were presented in the AB or BA order, randomly.

During the test, for each pair presentation, a participant was first presented with the two complete utterances and the target definition, and could freely listen to the performances A and B; the utterance was transcribed orthographically. Then, the participant switched to a screen where only the two versions (A and B) of a DM pair could be freely listened to (without the illocution). A third screen then presented the two complete utterances (i.e., the DM plus the illocution, in the two versions A and B) only once, one after the other, with a 500ms pause between them. After this final listening, participants had to select the DM that best fit the definition, clicking on the A or B button (the attribution of the two sentences to the A or B slots was done randomly). The next pair was then presented, following the same three-step procedure. Test completion took about 15 minutes.

8.6.2 Identification task

In the discrimination task, participants were presented the description of a DM function and had to decide which one of two prosodic realizations best match the function. In the identification task, on the other hand, participants were presented a stimulus and had to decide (respond) what function that prosodic realization best corresponds to.

The stimuli were presented with the prosodic characteristics of

one of the DMs modified in three versions. The prosodic forms will be referred to as **Descending, Ascending, Flat** – for short D/A/F. These references are preferred to CNT/EXP/INP, because using the functions' labels would introduce a confusion between *acoustic form* and function. Additionally, to the three audio forms (D/A/F), a Written (W) presentation modality was proposed.

During the test, for each stimulus, a participant was first presented with the audios of the DM and of the complete utterance. They could freely listen to the DM and the utterance, but they had to listened to both before being given the possibility to choose a function. When an audio was presented, participants were not given an orthographic transcription. This was done only for the written presentations in which case no audio was given. For written stimuli, participants were presented the DM and the illocutionary unit separated by a comma, in the form “uai, vamo ver” (well, let's see). After hearing the audios or having some time to read the written stimulus, participants were presented three boxes containing the description of the functions with the following content:

- Anunciar uma conclusão (Announce a **conclusion** – which should be matched with Descending/CNT – function coded as CON)
- Manifestar uma surpresa (Show **surprise** – which should be matched with Ascending/EXP – function coded as SUR)
- Apenas começar a frase (Simply **start** the phrase – which should be matched with Flat/INP – function coded as STA)

The selection was forced-choice. Each participant was presented 24

stimuli from the Latin square plus four stimuli with “uai, vamo ver” as a training start. Each group of the Latin square contained one version of the manipulated DMs for each lexeme. The stimuli were based on 24 utterances, each composed of a MD and their COM part. The DMs were based on seven (plus one more “ah”) different lexeme - *ah* (presented twice), *eh*, *gente*, *não*, *oh*, *pô*, *uai*, each being presented with three functions (CNT, EXP, INP), making up the 24 (8*3) stimuli, plus four training stimuli per participant. The next stimulus was then presented, following the same procedure. Test completion took about 10 minutes. The first stimuli used for training were not taken into account in the analysis. To sum up:

- a) The listeners were distributed in four groups, following a Latin square distribution, so each group was presented one of the 24 utterances once – with one modality Ascending (A), Descending (D), Flat (F) or Written (W);
- b) Each group was presented with a given utterance with a different modality;
- c) Each group saw all 24 utterances, and were presented with the same number of stimuli with a given modality, and to all the 7 lexemes.

8.7 ANALYSIS OF THE DISCRIMINATION TASK RESULTS

The findings presented in this section were first published in Raso et al. (to appear). The A or B answer to each pair was expressed as a “Match” if the selected DM’s prosodic characteristics actually matched the proposed definition (or as a “Miss” if not). There were therefore six types of pairs plus definition: in the following notation, the first DM of a pair (marked in bold) also corresponds to the presented definition (i.e., the boldface function corresponds to the presented definition, a

“Match” answer is thus equal to this boldface function). A listener was presented with the following set of Pairs: (**CNT**-EXP), (**CNT**-INP), (**EXP**-CNT), (**EXP**-INP), (**INP**-CNT), (**INP**-EXP). These six pairs of stimuli were presented through three lexical Contexts: the three sentences (the DM + illocution) – thus, 18 presentations to each participant. Each Pair was presented (randomly) in a given Order (AB or BA).

8.7.1 Binomial generalized model

These three factors, the presented Pair, the lexical Context, and the presentation Order, were used as fixed factors in a binomial generalized linear model to explain the variation in the proportion of (Match, Miss) answers (dependent variable) by the 85 participants.

Following Crawley (2012), a maximal model was fit (using the `glm()` function of the R software, R Core Team, 2022), with the dependent variable (proportion of Match answers) explained by the three fixed factors plus all their double and triple interactions. This maximal model was then submitted to a simplification process, removing iteratively the higher order interactions, when this did not lead to a significant loss of explanatory power in the model. The simplification steps are summarized in Table 45, which presents the model simplification process (output of R’s `step()` function), with the interactions or factors tested at each step, and the progressive reduction of the AIC criterion. The last row contains the minimal adequate model. The minimal adequate model contains only the Pair factor, that explains variations in the proportion of (Match, Miss) answers.

Table 45 - summary of the model simplification process (output of R’s `step()` function)

	Resid. df	Resid. deviance	Df	Deviance	Pr(>Chi)
<hr/>					

Start: Model: (Match, Miss) ~ Pair * Context * Order						
<none>		0	0			
- Pair:	10	6.921	-	-6.921	0.7329	
- Context:Order			10			
Step 2: Model: (Match, Miss) ~ (Pair + Context + Order)^2						
<none>	10	6.921				
- Pair: Context	20	21.038	-	-14.117	0.1677	
			10			
- Context:Order	12	11.303	-2	-4.382	0.1118	
- Pair:Order	15	15.600	-5	-8.679	0.1226	
Step 3: Model: (Match, Miss) ~ (Pair + Context) * Order						
<none>	20	21.038				
- Context:Order	22	25.638	-2	-4.600	0.1003	
- Pair:Order	25	30.112	-5	-9.074	0.1062	
Step 4: Model: (Match, Miss) ~ Pair + Context + Order + Pair:Order						
<none>	22	25.638				
- Pair:Order	27	34.532	-5	-8.894	0.1134	
- Context	24	31.409	-2	-5.771	0.0558	
Step 5: Model: (Match, Miss) ~ Pair + Context + Order						
<none>	27	34.532				
- Order	28	34.567	-1	-0.035	0.8510	
- Context	29	39.763	-2	-5.231	0.0731	
- Pair	32	133.034	-5	-98.503	< 2.2e-16	

Step 6: Model: (Match, Miss) ~ Pair + Context					
<none>	28	34.567			
- Context	30	39.811	-2	-5.244	0.0727
- Pair	33	133.047	-5	-98.480	< 2.2e-16
Step 7: Model: (Match, Miss) ~ Pair					
<none>	30	39.811			
- Pair	35	138.019	-5	-98.209	< 2.2e-16

The binomial regression, detailed above, showed that only the type of Pair had a significant effect on the proportion of (Match, Miss) answers, the other two independent variables (the lexical Context and the presentation Order, were dropped during the simplification phase). The model summary is proposed in Table 46, presenting the values of the binomial model's coefficients; the (CNT-EXP) level of the Pair factor was used for intercept.

Table 46 - Output of the minimal adequate model, presenting the values of the binomial model's coefficients; the (CNT-EXP) level of the Pair factor was used for intercept. Uncertainty intervals (profile-likelihood) and p-values (two-tailed) computed using a Wald z-distribution approximation

Parameter	Log-Odds	SE	95% CI	z	p
(Intercept)	0.80	0.14	[0.54, 1.07]	5.91	< .001
Pair (CNT-INP)	-0.47	0.19	[-0.83, -0.10]	-2.51	0.012
Pair (EXP-CNT)	0.88	0.22	[0.46, 1.32]	4.02	< .001
Pair (EXP-INP)	1.33	0.24	[0.86, 1.83]	5.45	< .001

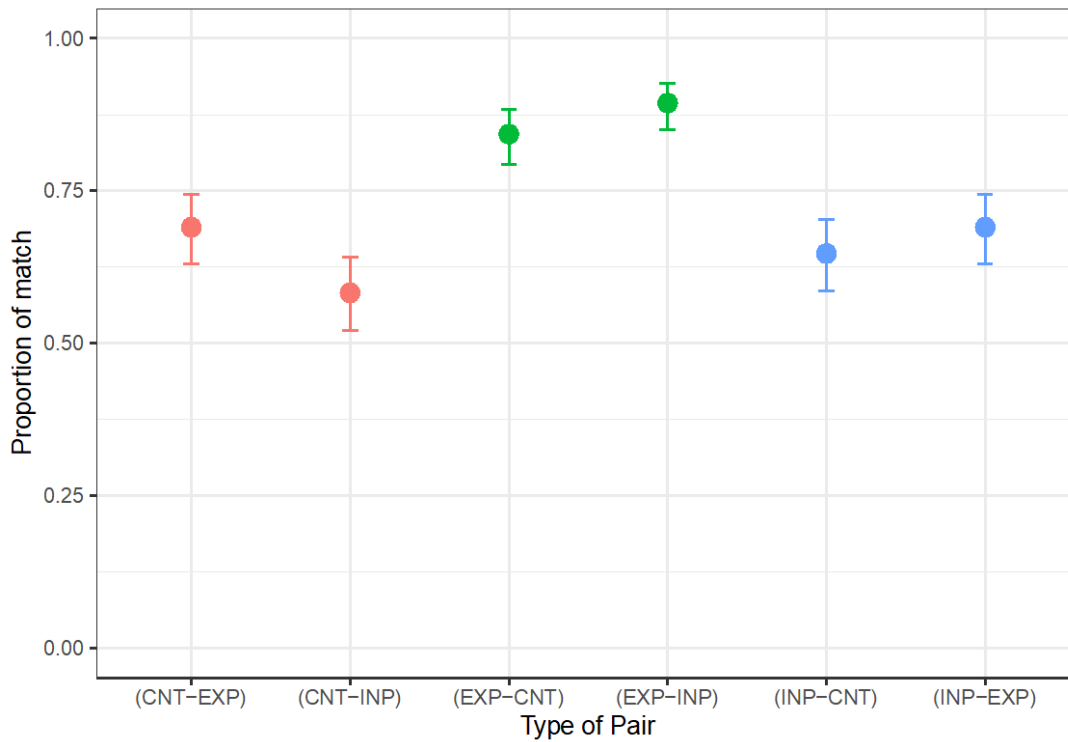
Pair (INP-CNT)	-0.19	0.19	[-0.57, 0.17]	-1.03	0.301
Pair (INP-EXP)	-3.04e-16	0.19	[-0.38, 0.38]	-1.59e-15	> .999

8.7.2 Proportion of Match by Pair

Figure 73 shows the proportion of Match predicted by the model for the six types of Pair + definition (Order not having a significant effect, the first part of each pair indicates the proposed definition). Four Pairs, (CNT-EXP), (CNT-INP), (INP-CNT), and (INP-EXP), reached a comparable level of discrimination (from 60 to 70%), that do not show significant differences between themselves (see Table 46 - Output of the minimal adequate model, presenting the values of the binomial model's coefficients; the (CNT-EXP) level of the Pair factor was used for intercept. Uncertainty intervals (profile-likelihood) and p-values (two-tailed) computed using a Wald z-distribution approximation); conversely, the (EXP-CNT) and (EXP-INP) Pairs showed a significant rise of Match answers compared to all the other Pairs (above 80% of Match).

Figure 73 - Proportion of Match answers fitted by the binomial

regression as a function of the type of Pair



8.8 ANALYSIS OF THE IDENTIFICATION TASK

Participants answered which of the three proposed function they thought best fit each stimulus. This categorical answer (CON/SUR/STA) is used as a dependent variable to a multinomial regression (fit using the `multinom()` function of R's "nnet" library; Venables & Ripley, 2010) – thus we observed the variation in the proportion of each category (CON/SUR/STA) in the participants' answer according to the following independent variables:

- a) The presentation **Modality** (four levels: D/A/F/W)
- b) The **Lexeme** used for the DM (seven levels:
ah/eh/gente/não/oh/po/uai)

- c) The functional **Class** of the DM in the utterance (three levels: CNT/EXP/INP)

8.8.1 Multinomial model

The model is expressed with the following formula, following Gries (2021) and using R's syntax:

$$\text{Answer} \sim 1 + \text{Modality} * \text{Lexeme} * \text{Class}$$

The model with three parameters was fitted to the proportion of answers observed in each possible function (Conclusion, Surprise, Start). The simplification of the model was tested, but removing the three-way interaction did lead to a significant loss in the model – which was thus kept. Table 47 presents the likelihood ratio tests of the Multinomial Models comparing the complete model to a model without the triple interaction:

Table 47 - Multinomial models - Complete model vs Model without triple factor interaction

Answer: AnswerID						
Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1 1 + (MODALITY + DM + CLASS)^2	11396	10221.815				
2 1 + MODALITY * DM * CLASS	11324	9963.957	1 vs 2	72	257.8577	0

The model's output is presented in Table 48 and through a series of figures that represent the mean tendencies estimated from the model for each of the factors and their interactions. The reference levels are – for Modality (W), for Lexeme (ah), for Class (CNT). The full model explains about 40% of the total variance ($R^2 = 0.414$).

Table 48 - Multinomial model's output - Identification task

	Estimate	Std. Error	z-values	P-values
(Intercept):SUR	-0.7674999	2.37E-01	-3.23E+00	1.22E-03
(Intercept):STA	-0.4420649	2.14E-01	-2.07E+00	3.85E-02
MODALITY.D:SUR	0.17313417	3.24E-01	5.35E-01	5.93E-01
MODALITY.D:STA	0.29348805	2.88E-01	1.02E+00	3.08E-01
MODALITY.A:SUR	1.0863096	3.32E-01	3.27E+00	1.07E-03
MODALITY.A:STA	0.76055551	3.16E-01	2.41E+00	1.60E-02
MODALITY.F:SUR	-0.8412311	4.54E-01	-1.85E+00	6.40E-02
MODALITY.F:STA	0.70430108	3.00E-01	2.35E+00	1.88E-02
LEXEME.eh:SUR	-12.1391	3.23E-01	-3.76E+01	0.00E+00
LEXEME.eh:STA	0.441934	3.59E-01	1.23E+00	2.18E-01
LEXEME.gente:SUR	1.54170112	4.22E-01	3.65E+00	2.60E-04
LEXEME.gente:STA	1.48437303	3.98E-01	3.73E+00	1.92E-04
LEXEME.não:SUR	-1.352839	4.93E-01	-2.74E+00	6.06E-03
LEXEME.não:STA	-0.9852789	3.86E-01	-2.55E+00	1.07E-02
LEXEME.oh:SUR	2.78205109	5.83E-01	4.77E+00	1.81E-06
LEXEME.oh:STA	1.94588392	5.93E-01	3.28E+00	1.02E-03
LEXEME.pô:SUR	1.79678677	4.38E-01	4.10E+00	4.12E-05
LEXEME.pô:STA	0.44177576	4.96E-01	8.91E-01	3.73E-01
LEXEME.uai:SUR	1.13459906	3.88E-01	2.93E+00	3.43E-03
LEXEME.uai:STA	0.72900894	3.78E-01	1.93E+00	5.37E-02
CLASS.EXP:SUR	0.55645749	3.31E-01	1.68E+00	9.27E-02
CLASS.EXP:STA	0.86372361	2.92E-01	2.96E+00	3.06E-03
CLASS.INP:SUR	-1.024674	3.92E-01	-2.61E+00	8.93E-03
CLASS.INP:STA	-0.2509021	2.95E-01	-8.49E-01	3.96E-01
MOD.D:LEX.eh:SUR	-7.0236511	3.89E-01	-1.81E+01	7.86E-73
MOD.D:LEX.eh:STA	-0.9293943	5.01E-01	-1.85E+00	6.37E-02
MOD.A:LEX.eh:SUR	10.6674262	3.23E-01	3.31E+01	8.12E-240
MOD.A:LEX.eh:STA	-1.5078001	5.15E-01	-2.93E+00	3.39E-03
MOD.F:LEX.eh:SUR	-4.6879704	3.92E-01	-1.20E+01	5.58E-33
MOD.F:LEX.eh:STA	-0.7039616	4.78E-01	-1.47E+00	1.41E-01
MOD.D:LEX.gente:SUR	-1.5343611	6.18E-01	-2.48E+00	1.31E-02
MOD.D:LEX.gente:STA	-1.2302136	5.49E-01	-2.24E+00	2.50E-02
MOD.A:LEX.gente:SUR	-0.9049054	6.09E-01	-1.49E+00	1.37E-01
MOD.A:LEX.gente:STA	-1.3319495	6.12E-01	-2.18E+00	2.96E-02

	Estimate	Std. Error	z-values	P-values
MOD.F:LEX.gente:SUR	0.22182027	6.95E-01	3.19E-01	7.50E-01
MOD.F:LEX.gente:STA	-0.4935599	5.57E-01	-8.87E-01	3.75E-01
MOD.D:LEX.não:SUR	-0.6551678	7.48E-01	-8.76E-01	3.81E-01
MOD.D:LEX.não:STA	-0.2158763	5.26E-01	-4.11E-01	6.81E-01
MOD.A:LEX.não:SUR	-0.98064	7.62E-01	-1.29E+00	1.98E-01
MOD.A:LEX.não:STA	-0.0951623	5.55E-01	-1.72E-01	8.64E-01
MOD.F:LEX.não:SUR	1.08982671	8.25E-01	1.32E+00	1.87E-01
MOD.F:LEX.não:STA	0.55595015	5.26E-01	1.06E+00	2.91E-01
MOD.D:LEX.oh:SUR	-3.1997803	7.47E-01	-4.28E+00	1.86E-05
MOD.D:LEX.oh:STA	-1.2505036	6.78E-01	-1.84E+00	6.53E-02
MOD.A:LEX.oh:SUR	-1.9375649	7.24E-01	-2.67E+00	7.48E-03
MOD.A:LEX.oh:STA	-1.1663417	7.34E-01	-1.59E+00	1.12E-01
MOD.F:LEX.oh:SUR	-0.991607	8.20E-01	-1.21E+00	2.27E-01
MOD.F:LEX.oh:STA	-1.2525982	7.31E-01	-1.71E+00	8.64E-02
MOD.D:LEX.pô:SUR	1.93313219	8.73E-01	2.21E+00	2.68E-02
MOD.D:LEX.pô:STA	0.40037483	1.02E+00	3.94E-01	6.94E-01
MOD.A:LEX.pô:SUR	1.38122055	8.73E-01	1.58E+00	1.13E-01
MOD.A:LEX.pô:STA	-12.460575	3.37E-01	-3.70E+01	4.38E-300
MOD.F:LEX.pô:SUR	16.2529971	4.26E-01	3.82E+01	0.00E+00
MOD.F:LEX.pô:STA	14.31567	4.26E-01	3.36E+01	4.47E-248
MOD.D:LEX.uai:SUR	1.81030513	6.87E-01	2.63E+00	8.44E-03
MOD.D:LEX.uai:STA	-1.2745262	9.64E-01	-1.32E+00	1.86E-01
MOD.A:LEX.uai:SUR	12.7595235	3.96E-01	3.22E+01	1.92E-227
MOD.A:LEX.uai:STA	11.1293792	3.96E-01	2.81E+01	1.55E-173
MOD.F:LEX.uai:SUR	3.87489255	9.04E-01	4.29E+00	1.80E-05
MOD.F:LEX.uai:STA	0.10665185	9.24E-01	1.15E-01	9.08E-01
MOD.D:CLA.EXP:SUR	-1.0973622	4.81E-01	-2.28E+00	2.26E-02
MOD.D:CLA.EXP:STA	-0.9119158	4.02E-01	-2.27E+00	2.34E-02
MOD.A:CLA.EXP:SUR	-0.5120142	4.65E-01	-1.10E+00	2.71E-01
MOD.A:CLA.EXP:STA	-1.5567355	4.65E-01	-3.35E+00	8.04E-04
MOD.F:CLA.EXP:SUR	0.86941925	5.93E-01	1.47E+00	1.42E-01
MOD.F:CLA.EXP:STA	0.02645984	4.29E-01	6.17E-02	9.51E-01
MOD.D:CLA.INP:SUR	-1.2715235	6.83E-01	-1.86E+00	6.25E-02
MOD.D:CLA.INP:STA	-0.6989976	4.24E-01	-1.65E+00	9.95E-02
MOD.A:CLA.INP:SUR	-0.4728121	5.12E-01	-9.24E-01	3.56E-01

	Estimate	Std. Error	z-values	P-values
MOD.A:CLA.INP:STA	-1.5335674	4.58E-01	-3.35E+00	8.10E-04
MOD.F:CLA.INP:SUR	-0.644471	7.50E-01	-8.59E-01	3.90E-01
MOD.F:CLA.INP:STA	-1.2737015	4.17E-01	-3.05E+00	2.28E-03
LEX.eh:CLA.EXP:SUR	10.6156875	4.33E-01	2.45E+01	5.68E-133
LEX.eh:CLA.EXP:STA	-1.9051096	5.30E-01	-3.59E+00	3.27E-04
LEX.gente:CLA.EXP:SUR	-1.1080397	5.86E-01	-1.89E+00	5.88E-02
LEX.gente:CLA.EXP:STA	-2.1941125	5.86E-01	-3.74E+00	1.82E-04
LEX.não:CLA.EXP:SUR	0.53364226	6.57E-01	8.12E-01	4.17E-01
LEX.não:CLA.EXP:STA	0.69675237	5.05E-01	1.38E+00	1.68E-01
LEX.oh:CLA.EXP:SUR	-1.3666603	7.08E-01	-1.93E+00	5.35E-02
LEX.oh:CLA.EXP:STA	-2.2132325	7.38E-01	-3.00E+00	2.73E-03
LEX.pô:CLA.EXP:SUR	-0.4220729	6.14E-01	-6.88E-01	4.92E-01
LEX.pô:CLA.EXP:STA	-0.863144	6.96E-01	-1.24E+00	2.15E-01
LEX.uai:CLA.EXP:SUR	1.27364951	6.94E-01	1.84E+00	6.64E-02
LEX.uai:CLA.EXP:STA	-0.4571967	7.46E-01	-6.12E-01	5.40E-01
LEX.eh:CLA.INP:SUR	12.405187	4.21E-01	2.95E+01	4.93E-191
LEX.eh:CLA.INP:STA	-1.2750308	5.41E-01	-2.36E+00	1.84E-02
LEX.gente:CLA.INP:SUR	1.80830946	6.53E-01	2.77E+00	5.64E-03
LEX.gente:CLA.INP:STA	-1.079613	7.01E-01	-1.54E+00	1.24E-01
LEX.não:CLA.INP:SUR	1.12968809	7.90E-01	1.43E+00	1.53E-01
LEX.não:CLA.INP:STA	0.76173036	5.54E-01	1.37E+00	1.69E-01
LEX.oh:CLA.INP:SUR	-1.7624074	7.47E-01	-2.36E+00	1.84E-02
LEX.oh:CLA.INP:STA	-0.9842286	6.79E-01	-1.45E+00	1.47E-01
LEX.pô:CLA.INP:SUR	0.74309379	6.09E-01	1.22E+00	2.23E-01
LEX.pô:CLA.INP:STA	-0.153651	6.53E-01	-2.35E-01	8.14E-01
LEX.uai:CLA.INP:SUR	-0.1309569	6.27E-01	-2.09E-01	8.35E-01
LEX.uai:CLA.INP:STA	-0.1318734	5.29E-01	-2.49E-01	8.03E-01
MOD.D:LEX.eh:CLA.EXP:SUR	-7.8919199	1.88E-06	-4.21E+06	0.00E+00
MOD.D:LEX.eh:CLA.EXP:STA	0.57456579	7.65E-01	7.51E-01	4.52E-01
MOD.A:LEX.eh:CLA.EXP:SUR	-11.404081	5.92E-01	-1.93E+01	9.55E-83
MOD.A:LEX.eh:CLA.EXP:STA	2.91485386	7.46E-01	3.90E+00	9.42E-05
MOD.F:LEX.eh:CLA.EXP:SUR	-10.471388	1.96E-06	-5.34E+06	0.00E+00
MOD.F:LEX.eh:CLA.EXP:STA	0.50372351	7.24E-01	6.95E-01	4.87E-01
MOD.D:LEX.gente:CLA.EXP:SUR	2.10211678	8.69E-01	2.42E+00	1.56E-02
MOD.D:LEX.gente:CLA.EXP:STA	2.4551159	7.96E-01	3.08E+00	2.05E-03

	Estimate	Std. Error	z-values	P-values
MOD.A:LEX.gente:CLA.EXP:SUR	1.49343504	8.45E-01	1.77E+00	7.72E-02
MOD.A:LEX.gente:CLA.EXP:STA	3.00298292	8.90E-01	3.37E+00	7.42E-04
MOD.F:LEX.gente:CLA.EXP:SUR	-0.6396237	9.08E-01	-7.05E-01	4.81E-01
MOD.F:LEX.gente:CLA.EXP:STA	-0.028942	7.94E-01	-3.65E-02	9.71E-01
MOD.D:LEX.não:CLA.EXP:SUR	0.90528762	1.06E+00	8.56E-01	3.92E-01
MOD.D:LEX.não:CLA.EXP:STA	0.7011499	7.16E-01	9.79E-01	3.28E-01
MOD.A:LEX.não:CLA.EXP:SUR	-1.1285794	1.17E+00	-9.67E-01	3.33E-01
MOD.A:LEX.não:CLA.EXP:STA	0.67812176	7.56E-01	8.97E-01	3.69E-01
MOD.F:LEX.não:CLA.EXP:SUR	-0.4929838	1.03E+00	-4.77E-01	6.33E-01
MOD.F:LEX.não:CLA.EXP:STA	-1.265535	7.18E-01	-1.76E+00	7.80E-02
MOD.D:LEX.oh:CLA.EXP:SUR	4.71144041	1.04E+00	4.52E+00	6.11E-06
MOD.D:LEX.oh:CLA.EXP:STA	4.11241159	9.83E-01	4.18E+00	2.88E-05
MOD.A:LEX.oh:CLA.EXP:SUR	2.17353243	1.01E+00	2.15E+00	3.16E-02
MOD.A:LEX.oh:CLA.EXP:STA	3.06116893	1.06E+00	2.88E+00	3.99E-03
MOD.F:LEX.oh:CLA.EXP:SUR	-0.2416419	1.05E+00	-2.31E-01	8.17E-01
MOD.F:LEX.oh:CLA.EXP:STA	1.21415742	9.50E-01	1.28E+00	2.01E-01
MOD.D:LEX.pô:CLA.EXP:SUR	-2.6244963	1.07E+00	-2.46E+00	1.38E-02
MOD.D:LEX.pô:CLA.EXP:STA	0.59243144	1.18E+00	5.03E-01	6.15E-01
MOD.A:LEX.pô:CLA.EXP:SUR	-1.5937316	1.06E+00	-1.51E+00	1.32E-01
MOD.A:LEX.pô:CLA.EXP:STA	13.7256775	5.22E-01	2.63E+01	3.33E-152
MOD.F:LEX.pô:CLA.EXP:SUR	-18.543526	6.38E-01	-2.91E+01	1.27E-185
MOD.F:LEX.pô:CLA.EXP:STA	-15.45223	6.15E-01	-2.51E+01	1.67E-139
MOD.D:LEX.uai:CLA.EXP:SUR	13.4801762	5.23E-01	2.58E+01	1.84E-146
MOD.D:LEX.uai:CLA.EXP:STA	16.7643823	5.23E-01	3.21E+01	2.15E-225
MOD.A:LEX.uai:CLA.EXP:SUR	0.18688439	4.77E-01	3.92E-01	6.95E-01
MOD.A:LEX.uai:CLA.EXP:STA	1.91870389	4.77E-01	4.02E+00	5.79E-05
MOD.F:LEX.uai:CLA.EXP:SUR	9.17556343	4.57E-01	2.01E+01	1.01E-89
MOD.F:LEX.uai:CLA.EXP:STA	12.1357541	4.57E-01	2.66E+01	1.77E-155
MOD.D:LEX.eh:CLA.INP:SUR	8.16678486	3.89E-01	2.10E+01	8.44E-98
MOD.D:LEX.eh:CLA.INP:STA	1.9672228	7.41E-01	2.66E+00	7.90E-03
MOD.A:LEX.eh:CLA.INP:SUR	-9.2693381	5.30E-01	-1.75E+01	1.53E-68
MOD.A:LEX.eh:CLA.INP:STA	2.82574832	8.52E-01	3.32E+00	9.10E-04
MOD.F:LEX.eh:CLA.INP:SUR	6.67041614	3.92E-01	1.70E+01	5.71E-65
MOD.F:LEX.eh:CLA.INP:STA	2.10638951	7.36E-01	2.86E+00	4.22E-03
MOD.D:LEX.gente:CLA.INP:SUR	1.27534787	9.98E-01	1.28E+00	2.01E-01

	Estimate	Std. Error	z-values	P-values
MOD.D:LEX.gente:CLA.INP:STA	2.36576577	8.84E-01	2.68E+00	7.44E-03
MOD.A:LEX.gente:CLA.INP:SUR	2.2603486	1.09E+00	2.07E+00	3.81E-02
MOD.A:LEX.gente:CLA.INP:STA	2.39246756	1.33E+00	1.79E+00	7.27E-02
MOD.F:LEX.gente:CLA.INP:SUR	0.49364038	1.07E+00	4.63E-01	6.43E-01
MOD.F:LEX.gente:CLA.INP:STA	1.82062744	9.21E-01	1.98E+00	4.81E-02
MOD.D:LEX.não:CLA.INP:SUR	2.99579988	1.15E+00	2.60E+00	9.29E-03
MOD.D:LEX.não:CLA.INP:STA	0.91900142	7.75E-01	1.19E+00	2.35E-01
MOD.A:LEX.não:CLA.INP:SUR	0.30365954	1.10E+00	2.77E-01	7.82E-01
MOD.A:LEX.não:CLA.INP:STA	0.55247188	7.92E-01	6.98E-01	4.85E-01
MOD.F:LEX.não:CLA.INP:SUR	1.78264142	1.19E+00	1.50E+00	1.34E-01
MOD.F:LEX.não:CLA.INP:STA	0.7868999	7.42E-01	1.06E+00	2.89E-01
MOD.D:LEX.oh:CLA.INP:SUR	2.36285205	1.25E+00	1.88E+00	5.97E-02
MOD.D:LEX.oh:CLA.INP:STA	0.75833397	8.49E-01	8.93E-01	3.72E-01
MOD.A:LEX.oh:CLA.INP:SUR	2.45269488	9.59E-01	2.56E+00	1.05E-02
MOD.A:LEX.oh:CLA.INP:STA	1.9217158	9.18E-01	2.09E+00	3.64E-02
MOD.F:LEX.oh:CLA.INP:SUR	1.64143282	1.16E+00	1.41E+00	1.58E-01
MOD.F:LEX.oh:CLA.INP:STA	1.61801819	8.66E-01	1.87E+00	6.18E-02
MOD.D:LEX.pô:CLA.INP:SUR	-2.275693	1.15E+00	-1.99E+00	4.70E-02
MOD.D:LEX.pô:CLA.INP:STA	0.47879523	1.16E+00	4.14E-01	6.79E-01
MOD.A:LEX.pô:CLA.INP:SUR	-1.2955056	1.07E+00	-1.21E+00	2.27E-01
MOD.A:LEX.pô:CLA.INP:STA	13.3504141	5.42E-01	2.46E+01	4.21E-134
MOD.F:LEX.pô:CLA.INP:SUR	-15.515113	7.40E-01	-2.10E+01	1.18E-97
MOD.F:LEX.pô:CLA.INP:STA	-12.178482	5.98E-01	-2.04E+01	3.91E-92
MOD.D:LEX.uai:CLA.INP:SUR	-1.3093373	1.07E+00	-1.23E+00	2.19E-01
MOD.D:LEX.uai:CLA.INP:STA	1.64248924	1.09E+00	1.51E+00	1.32E-01
MOD.A:LEX.uai:CLA.INP:SUR	-11.619451	5.31E-01	-2.19E+01	2.50E-106
MOD.A:LEX.uai:CLA.INP:STA	-10.394189	5.09E-01	-2.04E+01	1.44E-92
MOD.F:LEX.uai:CLA.INP:SUR	-3.2100308	1.23E+00	-2.60E+00	9.20E-03
MOD.F:LEX.uai:CLA.INP:STA	-0.3572834	1.07E+00	-3.33E-01	7.39E-01

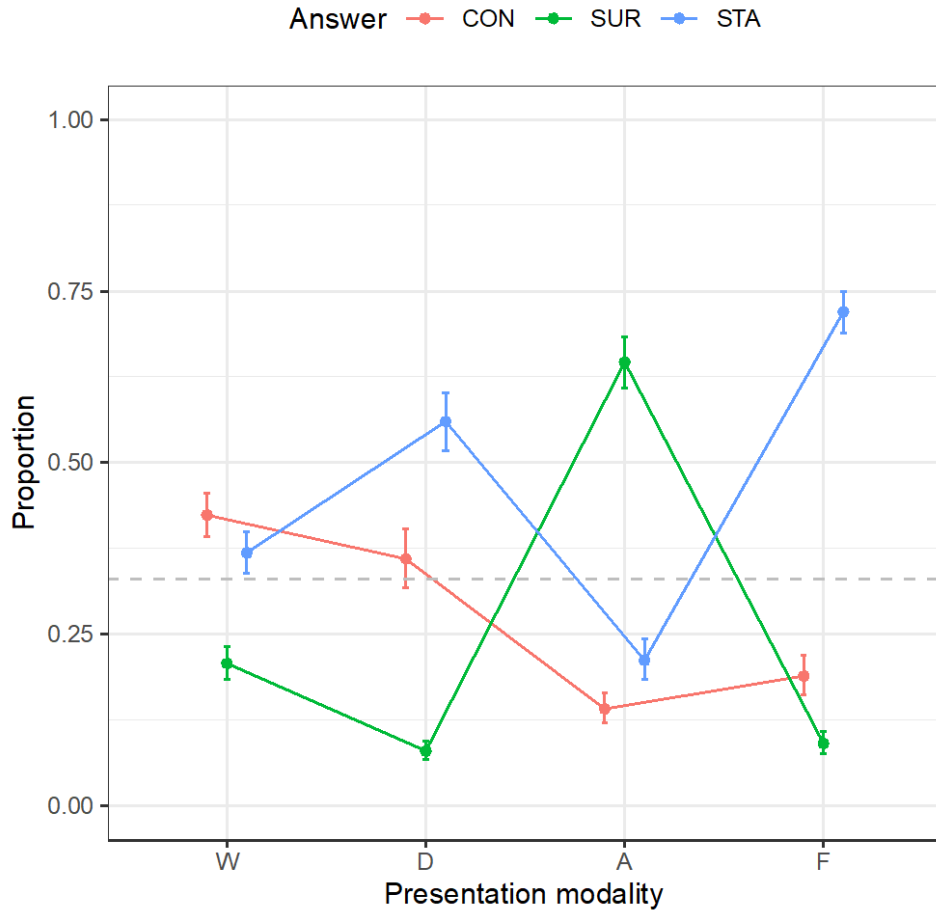
In the following subsections, the proportion of the three possible answers (Conclusion: CON; Surprise: SUR; Start: STA) are presented for the main effects of each factor and for their double and triple interactions.

8.8.2 Effect of the presentation Modality

Figure 74 shows the proportion of answers for each function (CON, SUR, STA) as a function of the presentation Modality – one of the three prosodic forms (D/A/F) or the written presentation (w). Our initial expectations were that the Descending (D) form would favor the Conclusion answer (CON); the Ascending (A) form would favor the Surprise (SUR) answer; the flat (F) would favor the Start (STA) answer. For the written (W) modality, we did not expect any particular result but it gives an idea of the bias linked to the lexical level; since here the effect of all utterance and lexemes is averaged, and because this effect is counterbalanced by construction, it shall give answers close to chance (1/3). In the figures, the grey dashed line stands for the proportion representing chance answer.

Figure 74 - Proportion of the (CON, SUR, STA) answers

for each level of the stimuli's presentation Modality



The first thing observable is that the W modality does not relevantly favor a functional interpretation. Answers are distributed around chance, but we can see that SUR is the least favored interpretation on the basis of the written presentations. On the other hand, by looking at the three modalities, we can see that the SUR interpretation is strongly impacted by prosody (being selected over 60% of the times for the Ascending contours, and not for the two other contours), and that this impact happens in accordance with initial expectations: A favors SUR. Also in line with initial expectations are the answers for the flat form stimuli: F favors STA. In its turn, the Descending form (D) disfavors SUR, but contrary to our expectations, the most favored

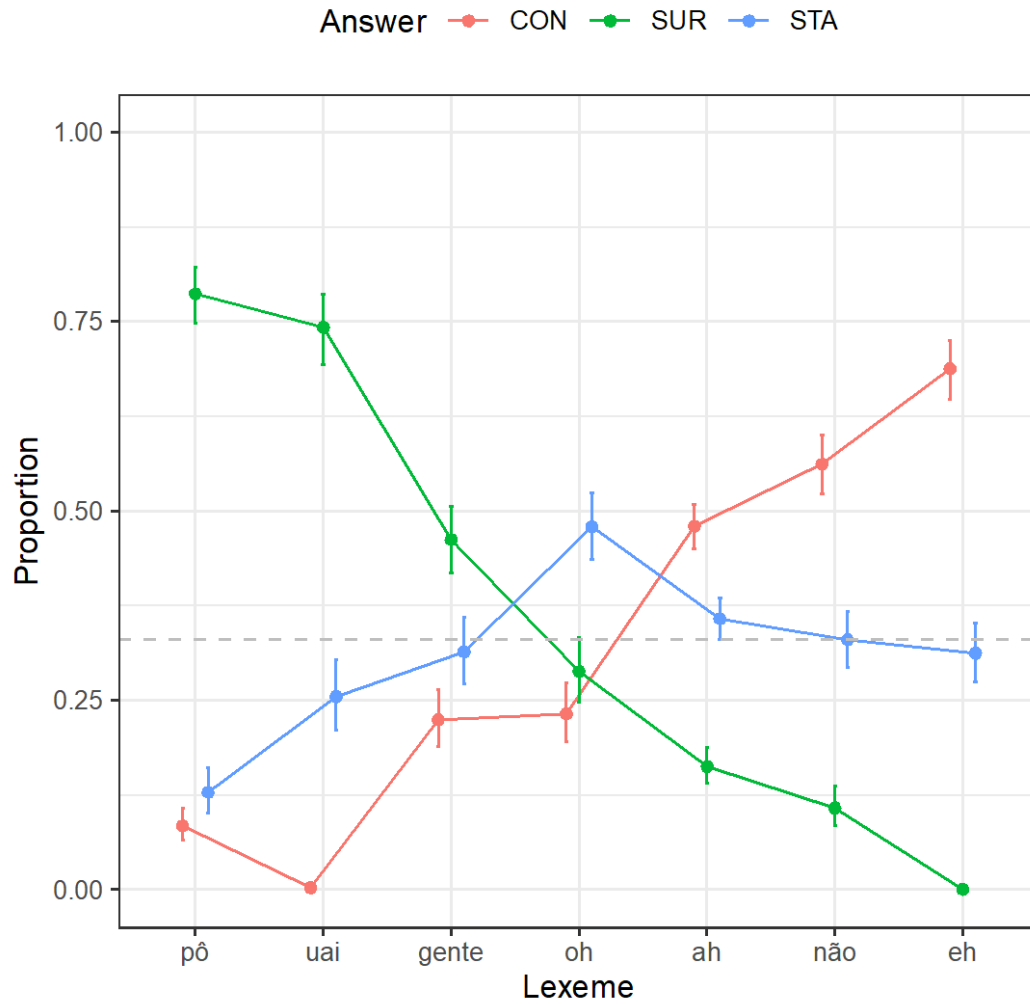
interpretation was STA but not CON – albeit D does not disfavor CON, as the F presentation does. Overall, we can say that (a) the interpretation of SUR seems to be the most directly dependent of prosody form both because it is disfavored by the written modality and because only one prosodic form favors its interpretation; and (b) that both D and F favors STA, with F the form most clearly linked to STA. However, we need to factor the lexeme in to check how prosody and the lexicon affected the functional interpretations. Before doing that, we check how the lexicon alone conditions the functional interpretation.

8.8.3 The lexeme

Figure 75 exhibits the proportion of answers as a function of the Lexeme. It was expected that if prosody only were in play, it would supersede the effect of the lexicon. Since factors are blocked, this would be translated into proportion of answers near the chance dashed line. However, we have some clear tendencies in the opposite direction.

Figure 75 - Proportion of the (CON, SUR, STA) answers for each level of

Lexeme used for the stimuli



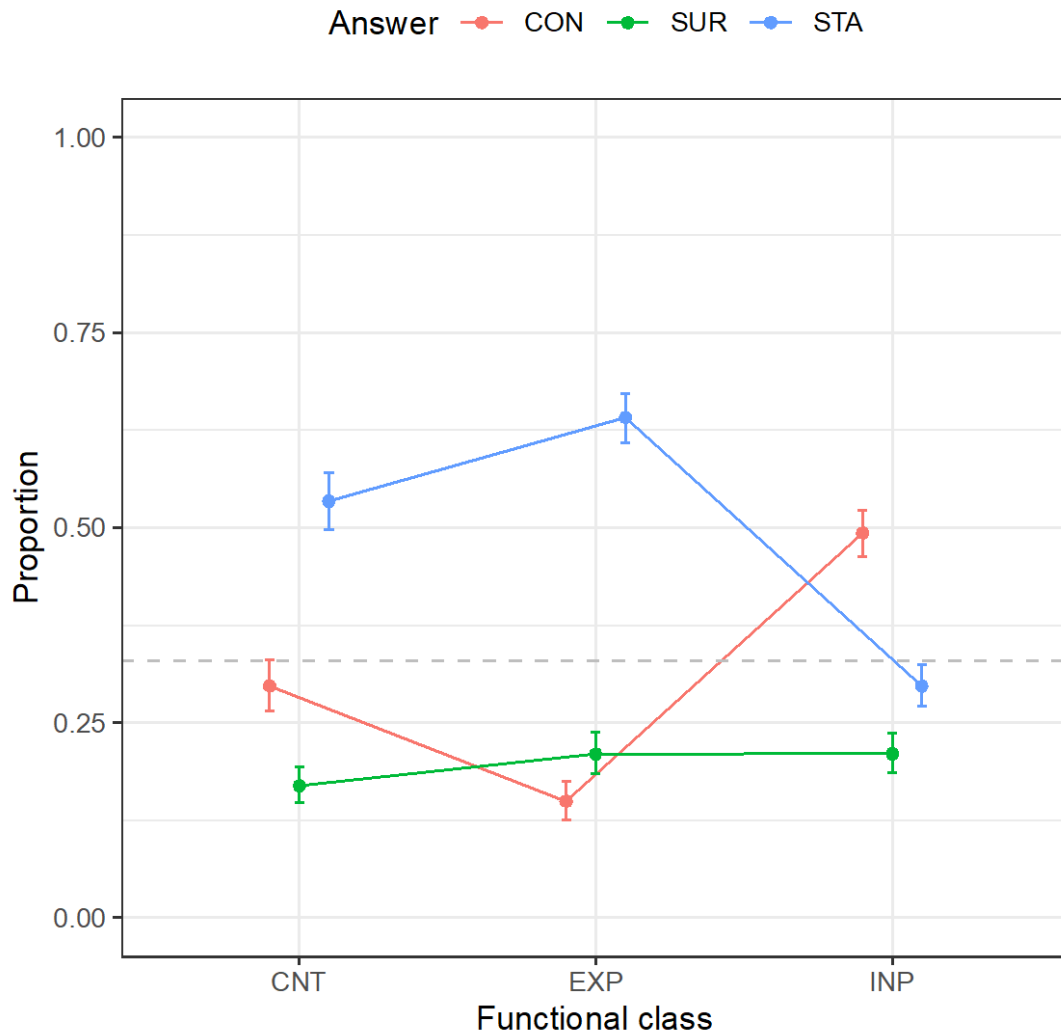
Firstly, the lexemes *pô* and *uai* clearly disfavor the functional interpretations of start and conclusion, while favoring surprise. On the other hand, *não* and *eh* disfavor SUR, favors CON, and are neutral to STA. Although showing some variation, *gente*, *oh* and *ah* seem to influence less the functional answer.

8.8.4 The functional CLASS attributed to the original stimuli

Figure 76 shows the proportion of answers as a function of the functional class originally assigned to each utterance selected to create the stimuli. As a reminder, each original utterance (DM plus illocution) was manipulated into three prosodic versions. An original CNT plus illocution would thus result in three DM, each with one function (CNT, EXP, and INT) plus the illocution. In this example, the original DM class was CNT. Expected results would, thus, show a random distribution across classes, if prosody only were playing a role in the functional attribution by participants. A result based on the original functional category would be shown by a strong match between the pairs CNT-CON, EXP-SUR, and INP-STA. This is not what happens. However, original CNTs and EXP seems to favor STA whereas INP favors CON answer.

Figure 76 - Proportion of the (CON, SUR, STA) answer for each level of

the functional Class of the stimuli

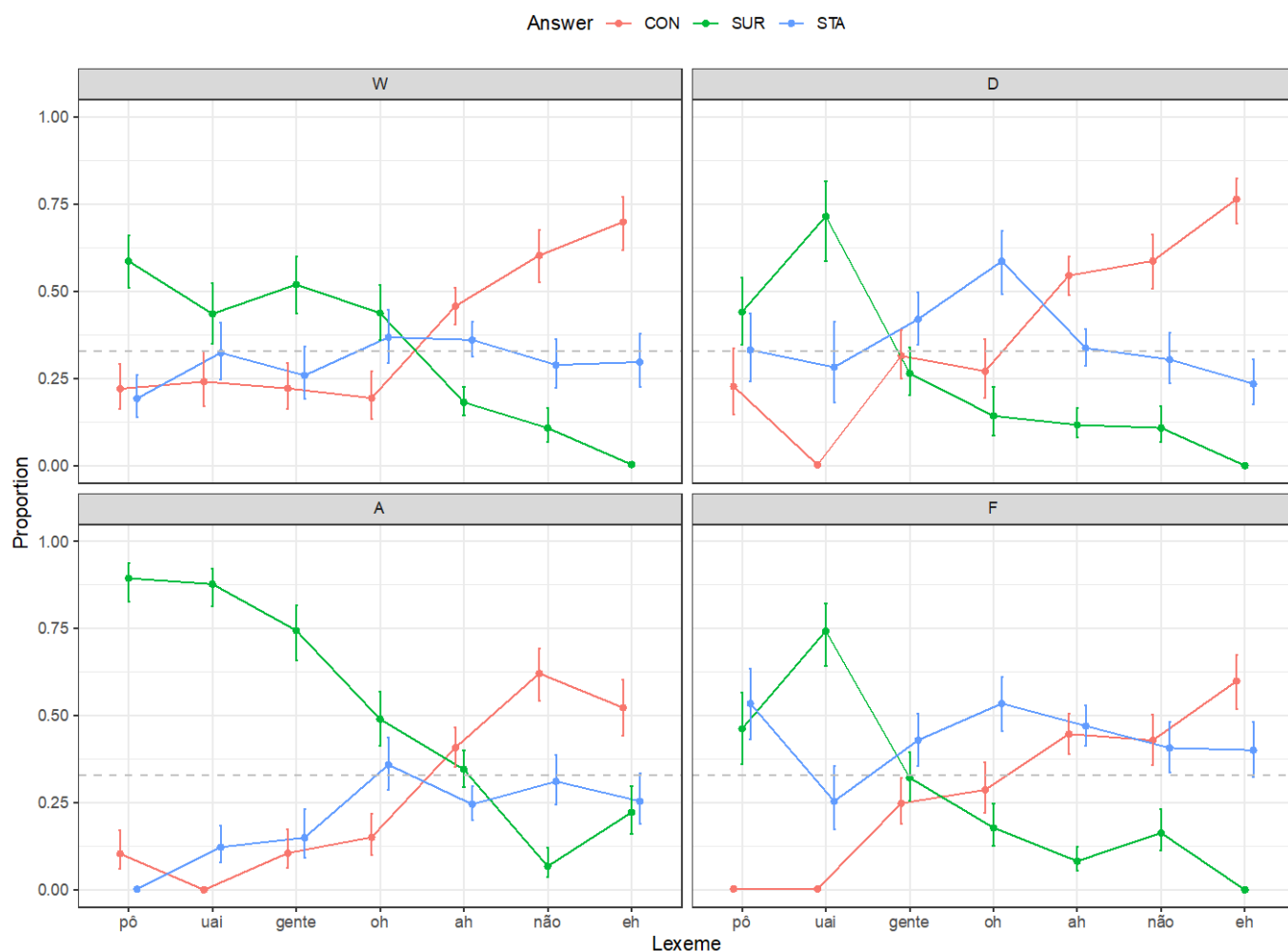


I will now present the results of the interactions between these three factors.

8.8.5 Interaction between presentation modality and lexeme

Figure 77 exhibits the interaction between the presentation modality (written or manipulated prosodic forms) and the lexeme.

Figure 77 - Interaction between modality and lexeme



For the written modality, there is a tendency for the lexemes *pô*, *uai*, *gente* to favor the SUR answer. The lexemes *ah*, *não* and *eh*, on the other hand, disfavor SUR and favors CON. Generally, the STA answer is not linked to any lexeme; only *pô* seem to disfavor STA to some extent.

The descending (D) form will affect negatively SUR answers in all lexemes but in *pô* and, especially, *uai*. Comparing W and D modalities, the SUR answers in surplus for the D presentations for *uai*

seems to come mostly from the CON answers observed in the W modality, while STA levels do not change. As a matter of fact, no matter what prosodic form *uai* takes on, it will tendentially be interpreted as a surprise. For the other Lexemes in the D modality, *ah*, *não* and *eh* favor the CON answers, whereas only *oh* will favor STA.

The ascending (A) form strongly affects the lexemes *pô*, *uai*, *gente*, and *oh* towards a SUR answers. Although this prosodic form is perceptually salient, the lexemes *não* and *eh* with A contours received a majority of CON answers. The STA answers are mostly disfavored for lexemes *pô*, *uai*, and *gente*.

Interestingly, the flat (F) form does especially favor, compared to other conditions, a STA answer for a given Lexeme. This is line with our expectations, but may be reinterpreted here as a double effect of prosody and Lexeme: the function of prosodic changes is primarily interpreted by participants under the bias of their semantic interpretation of the Lexeme. In the case of Flat prosodic form, only *uai* favors SUR and *eh* favors CON. Besides, flat *pô* and *uai* never elicit CON answers, and flat *eh* never elicit SUR answer: some interpretations of prosodic forms seem to be limited by the Lexeme.

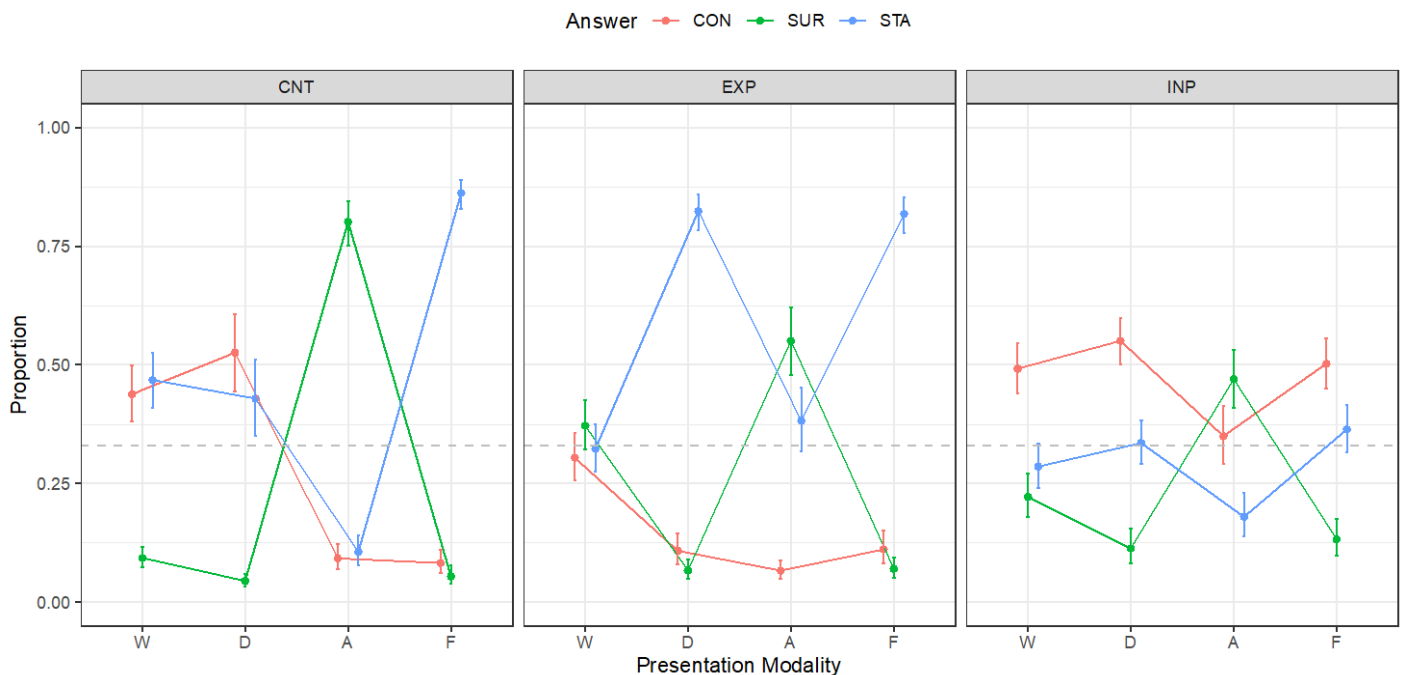
8.8.6 Interaction between modality and DM class

Figure 78 shows the proportion of answers as function of the original Class and the Modality (W/D/A/F). Expected results would higher proportions for pairs Descending-CON, Ascending-SUR, and Flat-STA, whatever the original Class of the utterance; this would indicate that the original class does not bias interpretation of the prosodic form.

Original CNT utterances with A prosody bias answers toward more SUR answers and F prosody bias towards STA answers – disfavoring the two other answers. In the case of D prosody, the CON answers are favored, while the SUR is disfavored but the STA answer are not disfavored. This is mostly in line with our expectations; let's note for this Class, the SUR answers are disfavored in the written presentations.

Original EXP utterances will favor the STA answer both with the Descending and the Flat forms. Only SUR is favored when we have the Ascending form. Here, the written modality does not elicit any particular functional interpretation (proportion around 1/3 for all functions). The CON answer is disfavored in all prosodic form. Here, it is possible to see that there seems to be an incompatibility between the illocutions typically introduced by CNT and EXP, since manipulating CNTs into EXPs did not pose any particular problem.

Figure 78 - Proportion of the (CON, SUR, STA) answers for each level of the presentation Modality of the stimuli for each functional Class



The original utterances with the INP Class favored the CON answer for written presentation with two of the three prosodic modalities (D and F) but not the Ascending modality – in that case, the prosody favors the SUR answers. So, it seems these illocutions favor a conclusive interpretation.

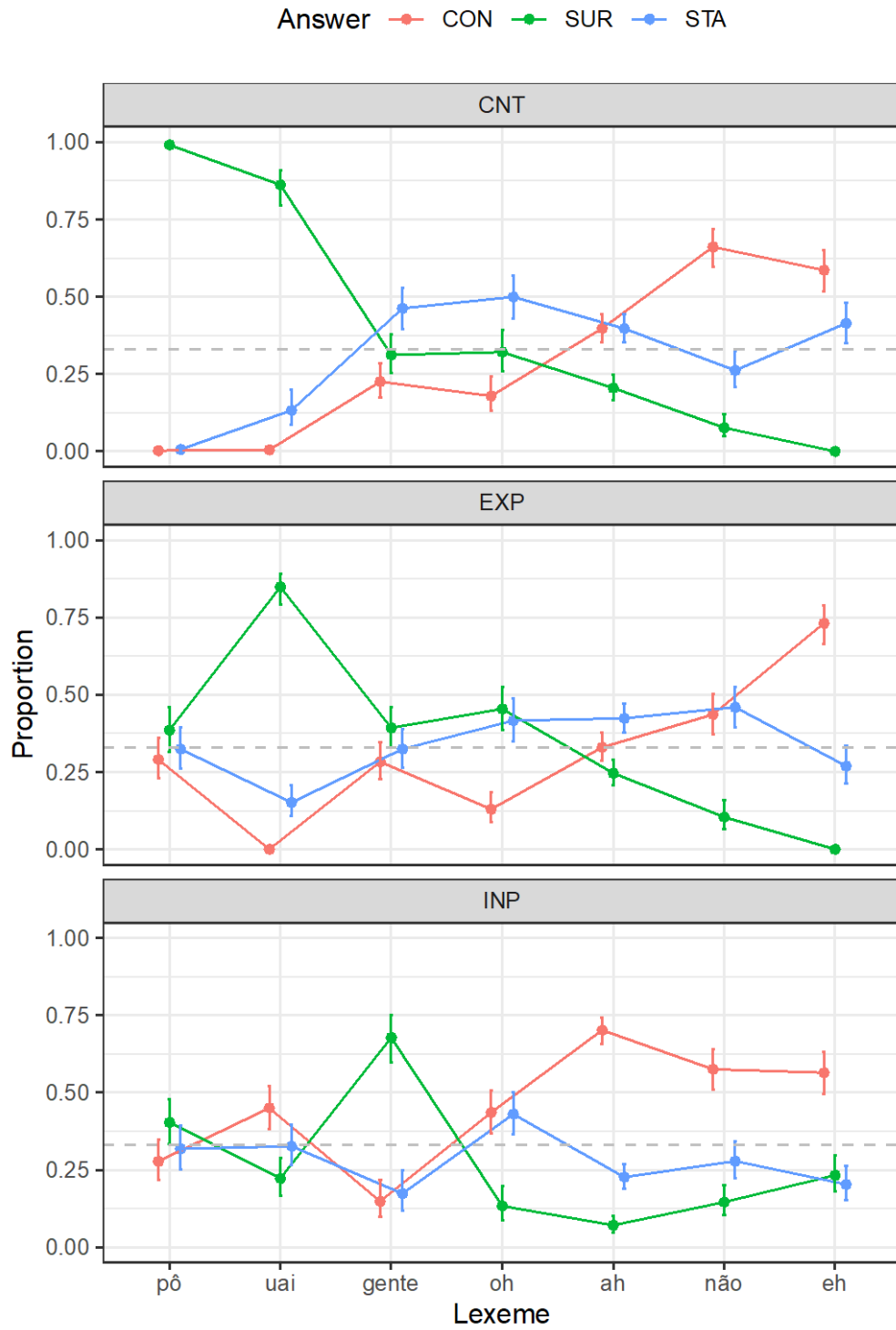
8.8.7 Interaction between lexeme and DM class

Figure 79 shows the effect on the proportion of answers of the interaction between the original DM Class and the Lexeme (averaging the effect of the presentation Modality). Considering that neither the context/illocution nor the lexeme influence functional answer would result in random distribution of answers, which is not the case.

Utterances classified originally as CNT favor SUR answers for the lexemes *pô* and *uai*, while *não* and *eh* favor the CON interpretation. *Gente*, *oh*, and *ah* present a more balanced distribution, although STA is favored. In original EXP utterances, *pô*, *gente*, and *ah* display similar distributions, whereas SUR is favored by the lexeme *uai* and disfavored by the lexemes *não* and *eh*, this last lexeme leaning towards CON once again. Original INP utterances do not strongly affect *pô* and *uai*. On the other hand, *gente* favors the SUR answer while *ah*, *não*, and *eh* favors the CON interpretation.

Figure 79 - Proportion of the (CON, SUR, STA) answers for

each level Lexeme and functional Class



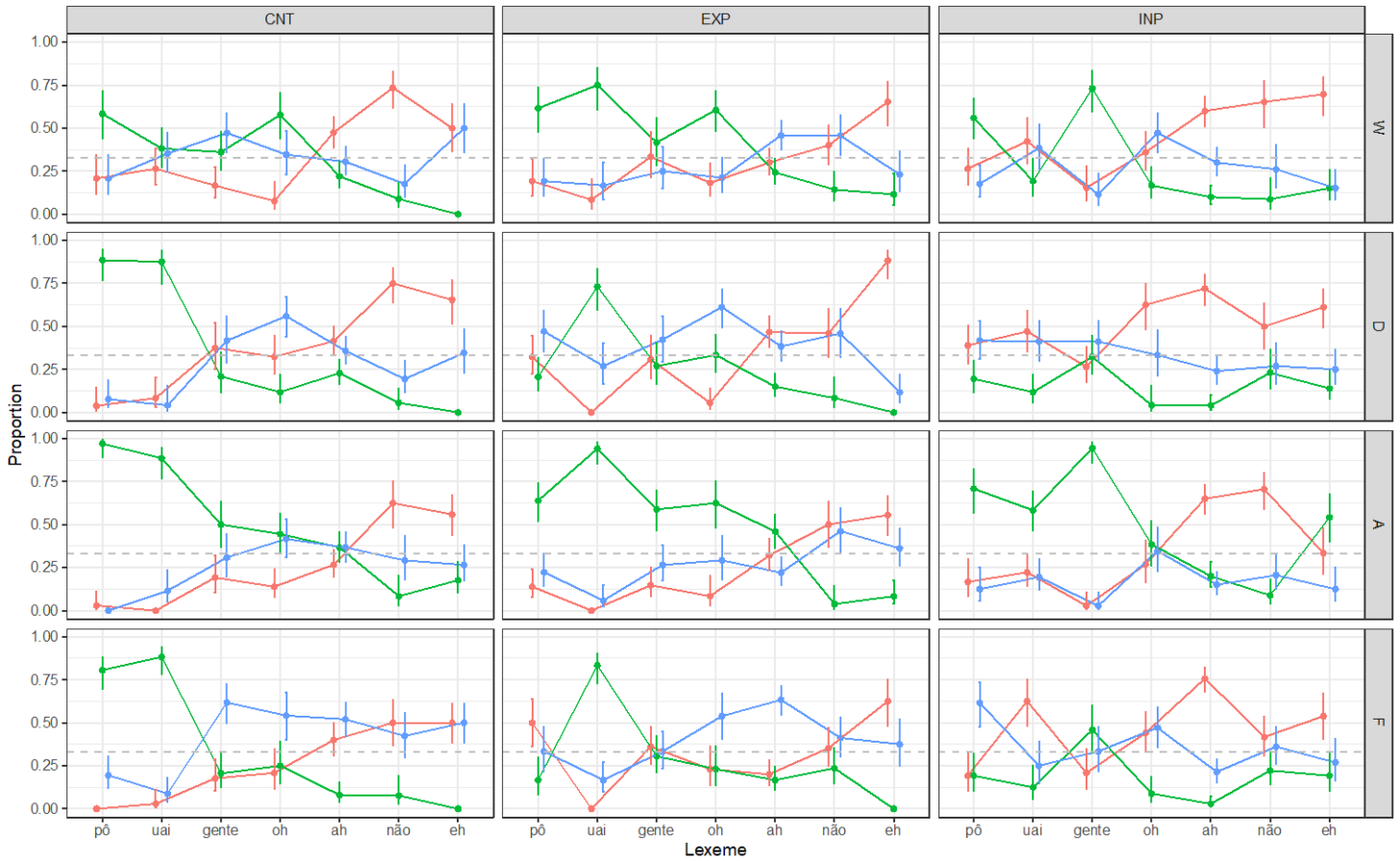
8.8.8 Triple interaction between MODALITY * LEXEME * CLASS

Figure 80 exhibits the triple interaction between presentation modality, original functional class, and lexeme. The surprise (SUR) answer is favored in most cases by the Ascending form and by the lexemes *pô*, *uai*, and *gente*. This answer is especially disfavored by the lexemes *não* and *eh*, and when the original utterance carried a CNT or an EXP. On the other extreme, the conclusive (CON) answer is favored by the lexemes *ah*, *não*, and *eh*, in almost combination of Modalities and original Class. However, the descending form (D), expected for CNT, tends to reinforce the CON answer in such lexemes, while the Flat form tends to reinforce the STA answer, often on par with CON on those lexemes. The STA answer, expected for utterance with Class INP, is favored by the lexeme *oh* and by descending and flat prosodic forms, especially with original CNT and EXP utterances.

Figure 80 - Proportion of the (CON, SUR, STA) answers for each level of

Lexeme, functional Class, and Presentation Modality

Answer — CON — SUR — STA



8.9 DISCUSSION OF RESULTS

The results of the discrimination test support the hypothesis of the importance of the prosodic realization in the functional interpretation of DMs in dialogic interactions. This first experiment was, however, more limited in terms of lexeme and utterances. It also showed that a discrimination paradigm allows the participants to better focus on prosodic form than a discrimination protocol, that favors more holistic interpretations of a given stimulus. A reduced set of three utterances and two lexemes ("ah", "gente") were tested. Albeit no effect of lexical Context was observed, more studies on a potential role of the other

linguistic levels were required. Another limitation was linked with the INP definition, that is not fully comparable to the two others definitions, being simpler. As shown during the preceding chapter, the INP function is also frequently observed with another prototypical contour (similarly flat and short, but with a much higher F0), which marks an attitude of contrast with what was said. However, this first perceptual validation has its strengths. First because it is based on fully spontaneous occurrences of DMs, a feature rarely observed in the perceptual evaluation of prosody, and that avoid the construction of artificial, unaccounted linguistic structures – and carries the original performance of a speakers in its complexity. Moreover, the prosodic characteristics of the DM part in the stimuli were derived from the theoretical description of the DMs' functions – thus no stimulus was exempt of quality bias linked to the resynthesis process, and all the DM carried an equivalent prosodic meaning, potentially removing features linked to affective or idiosyncratic characteristics of the original recordings. The paradigm used for this first evaluation was based on AB pair discriminations. This approach was preferred as the task was thought to be potentially complex, and pair comparisons allow an enhanced perception of subtle differences. The results, which clearly support the ability of listeners to select a prosodic form consistently for an association with a functional category, pleaded for validating the relationship of prosodic shape and functional definition using an identification task. A drawback of pair comparison is the rapid inflation in the number of presentations, each individual stimuli being paired with all the others: this strongly limits the ability to test many variations across factors, something an identification paradigm is better designed for.

Considering the limitations of the first test, a more complete evaluation for the same functions was conducted. This time, an identification paradigm was adopted, and a more complete set of lexemes was used. All examples were attested in the C-ORAL-BRASIL-I corpus, each one carrying, after modification, the three prosodic forms associated with the functional roles. Instead of two, a total of seven lexemes were tested (which is the maximal number of lexemes attested for these three functions in the corpus available). It seems that many factors are likely to have an effect and contribute to the final

interpretation of DMs' functional roles: the prosody of the DM, the lexeme, the type of illocution, and the context. This last factor is absent of the stimuli presented, and so participants may tend to reconstruct one so to semantically interpret the proposed utterances.

9 CONCLUSION

Discourse Markers were defined in this work as information units with interactional functions. As such, DMs are vehiculated through prosodic units, they have specific macro-functions that are conveyed by (or associated with) a prosodic form, and they have distributional preferences. DMs are not compositional with respect to the hosting pattern. Instead, they are aimed at regulating interactional aspects of the discourse. They may promote social cohesion (ALL), draw the addressee's attention to an illocutionary solution (CNT), express surprise without illocutionary force (EXP), highlight a previous content (EVD/HGL), or simply begin the utterance (INP). Each proposed function can be filled with a varied range of lexemes or small expressions, as shown in Table 12 - Lexical frequency by DM class. The lexicon has been shown to be variable whereas the prosodic form can account for the recognition of the proposed functions with good performance in a classification task. It has also been shown that, overall, prosody has positive effects on the recognition of DM functions. Nonetheless, the lexicon has been also shown to play an important role in the interpretation of DMs' role.

A classification model was presented together with the most relevant features for the distinction of each DM class against the others. It is possible to say that the classification model presents a good performance (accuracy scores varying between 68% and 78% for five classes). This model does not present the same accuracy level as those presented in Gobbo (2019) – around 80%. However, the current proposal (and respective model) is more complex: the task was carried out not with three but five DM classes. Moreover, the current model accounts for previously unclassified observations that were left out either because they were ambiguous or did not fit any existing class. Finally, the model was evaluated with more robust techniques, and its performance may reflect more reliably what happens in the wild. Another interesting observation concerns the features most frequently chosen by the one-vs-others models. In most cases, features involving fundamental frequency were important. Exceptions are the ALL class, which selects duration, and INP, which selects intensity and duration.

Alignment features also proved to be relevant for the distinctions.

An interesting approach for a more comprehensible model could be to have feature-dedicated models. An ensemble could be built that congregates models, each dedicated to a Discourse Markers facet. For instance, a model could vote based exclusively on the prosodic features. Another one could be in charge of aspects related to the distribution of the DM within the pattern. Not only a categorical feature indicating position (initial/medial/final) could prove useful, but also more fine-grained features could be tested that reflect the relative distance of the DM with respect to the illocutionary unit (both in relative time unit and of the number of information units), as well as neighboring information units. Another model could be responsible for judging the class based on the DM's lexical filling. This could be achieved using sentence embeddings (sentence transformers – Reimers & Gurevych, 2020) as input. This would prevent, for instance, first names (frequently used in CNT and INP) from being dealt with as very divergent categories, like a simple categorical encoding of the DM's text.

Furthermore, some important conclusions can be drawn from the perceptual tests. Firstly, some lexemes received strong associations with some functions. These associations seem to be favored or disfavored by prosody (and by the illocution). For example, *não* (no) has a conclusive function, but the perception of this Lexeme as conclusive becomes more salient when the prosodic form is descending and less salient when ascending. But the near categorical attribution of functionality by the test participants to some lexemes (e.g., *uai* or *eh*) shall not be overstated: first it may vary a lot according to the contextual interpretation of the lexeme (no lexeme has a fixed interpretation from the written form, across all utterances), and second – if participants rely on the basic meaning of lexemes, the corpus observations shown these lexemes are used in a variety of contexts. It may be that the participants struggle with the desemanticized nature of DM in this case. More works will be required to offer experimental protocols able to cope with this limitation (an association protocol, as in Shochi et al., 2020, may prove interesting). However, the lexicon offers a large set of possibilities, especially since proper names can be used as DMs in CNT and ALL. Moreover, the lexicon is pluri-functional; the

meaning of lexemes is highly dependent on the context. This makes a functional classification from lexicon complicated. Furthermore, the lexicon is very variable diastatically, diaphasically and diachronically, making a classification even more complicated. Secondly, both the type of illocution and its semantic content have been shown to have an effect on the global interpretation of the utterance. Of course, what was asked was the interpretation of the DM, which is embedded in the utterance, but the whole structure is producing a global meaning. It is reasonable to think that the DM functional role is holistically interpreted within the utterance. Therefore, if the illocution carries (or is interpreted³⁰ as) an illocutionary surprise, a conclusion, or something else, that affects the participant's interpretation of the DM. However, the function of the DM is, to a large extent, independent of the illocution. It seems sensible to think that the illocution imposes some combinatorial constraint on the DM, but there must be a degree of freedom. In addition, there are also many illocution categories (most of which are still in need of deeper descriptions), and this factor cannot be controlled for. What was taken into account was a simplified original DM class factor (that should sum up all the natural characteristics of the original utterance).

This research identified five prosodic forms that seem functionally coherent and sufficient to cover all the functions of the DMs. Prosody can vary diastatically and diaphasically (perhaps diachronically too) based on attitudinal parameters: higher or lower intensity, f0 range, articulation rate can undoubtedly depend on the demographic characteristics of participants (gender, age, socio-cultural level, and others) and on the situation (people adjust their attitude depending on the communicative situation). But the prosodic form (movement and alignment, first and foremost) remains, to a great extent, constant. This constancy is exactly what allowed for the good

³⁰ Let's note that in the case of the stimuli used for this experiment, their interactional interpretation is really difficult without having access to the history of the dialogues they were excised from. And, as shown by the results of the factor Class, presented in its written Modality, the participants did attribute some functions to the sentences that were not in line with the original categories.

classification scores obtained.

Based on these considerations, the provisional conclusion drawn from the experiments is that to categorize DMs, given that many factors influence their interpretation, one should start with the less variable factor –the prosodic form, knowing that other factors can modify the basic interpretation of the form and can even modify it a lot. One should not, advisably, start with the factors that vary the most, such as the lexicon, attitudes, or types of illocutions. These are all factors with important degrees of variation that do not allow for an initial organization. If we, by way of example, start with the lexicon, we will come to the conclusion that the same lexeme can accomplish essentially different functions and that the same function can be accomplished by entirely different lexemes, say, a proper name and a verb.

As it was argued, the prosodically-based proposed DMs are macro-functions that can take on more specific subfunctions depending on the context. However, the subfunctions are coherent with the macro-functions. For instance, CNT is thought to point to the illocutionary solution, i.e., to point to the intention of the speaker. If the speaker says something, then interrupts themselves to introduce a new planification, this repair can be introduced by a CNT (if pointing to a conclusion) or by an INP, if the speaker wants, for instance, mark a strong contrast with the interrupted ideation.

Finally, the design of the experiments showed some limitations: the wording of the questions, the nature of the data, the definition of the possible categories used to answer are notably complex and may not allow a smooth understanding by some participants. Naïve speakers are not taught during schooling to identify prosody or DMs the same way they are taught to interpret a lexeme like *uai* as an interjection that can express surprise. Here, the written bias, inherited from the educational system, may have played a significant role in our results. Reflection is therefore necessary to help design other experiments that takes into account the issues observed for metalinguistic tasks presented to naïve participants. A possible idea is to present only natural examples, without manipulations and

decontextualizations, and ask participants to identify the function.

10 REFERENCES

- A. Viterbi. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2), 260–269. <https://doi.org/10.1109/TIT.1967.1054010>
- Abraham, W. (1991). The grammaticization of the German modal particles. *Approaches to grammaticalization*, 2, 331–380.
- Ba, H., Yang, N., Demirkol, I., & Heinzelman, W. (2012). *BaNa: A Hybrid Approach for Noise Resilient Pitch Detection*. <https://doi.org/10.13140/2.1.3921.6321>
- Bagshaw, P. C., Hiller, S. M., & Jack, M. A. (1993). Enhanced Pitch Tracking and the Processing of F0 Contours for Computer Aided Intonation Teaching. *Proc. Eurospeech '93, Berlin, 2*, 1003–1006.
- Bally, C. (1950). *Linguistique générale et linguistique française* (3^o ed). A. Francke.
- Barbosa, P. (2013). Semi-automatic and automatic tools for generating prosodic descriptors for prosody research. Em B. Bigi & D. Hirst (Orgs.), *Proceedings* (p. 86–90).
- Barbosa, P. A. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication*, 49(9), 725–742. <https://doi.org/10.1016/j.specom.2007.04.013>
- Bartošek, J. (2011). A Pitch Detection Algorithm for Continuous Speech Signals Using Viterbi Traceback with Temporal Forgetting. *Acta Polytechnica*, 51(5). <https://doi.org/10.14311/1422>
- Batra, M. D., JAYESH, & Ramalingam, C. S. (2022). Robust Pitch Estimation Using Multi-Branch CNN-LSTM and 1-Norm LP Residual. *Proc. Interspeech 2022*, 3573–3577. <https://doi.org/10.21437/Interspeech.2022-10704>
- Bazzanella, C. (1990). Phatic connectives as interactional cues in contemporary spoken Italian. *Journal of Pragmatics*, 14(4), 629–647. [https://doi.org/10.1016/0378-2166\(90\)90034-B](https://doi.org/10.1016/0378-2166(90)90034-B)
- Bechtold, B. (2021). *Pitch of Voiced Speech in the Short-Time Fourier Transform: Algorithms, Ground Truths, and Evaluation Methods*. Universität Oldenburg.

- Bellman, R., Corporation, R., & Collection, K. M. R. (1957). *Dynamic Programming*. Princeton University Press. <https://books.google.com.br/books?id=wdtoPwAACAAJ>
- Bick, E., Mello, H., Panunzi, A., & Raso, T. (2012). The annotation of the CORAL-BRASIL spoken corpus using an adaptation of the Palavras Parser. In: CALZOLARI, N. et al. *Proceedings of LREC, 2012*, 3382–3386.
- Biq, Y.-O. (1990). Conversation, continuation, and connectives. *Text-Interdisciplinary Journal for the Study of Discourse*, 10(3), 187–208.
- Blanche-Benveniste, C. (1997). The unit in written and oral language. Pontecirvo C., (éd.), *Writing Development. An interdisciplinary view.*, John Benjamins Publishing Company, Amsterdam/Philadelphia, 21–45.
- Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings 17*, 97–110.
- Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer [Computer program]. Version 6.2.08*. <http://www.praat.org/>
- Bois, D., W, J., Chafe, W. L., Meyer, Ch., Thompson, S. A., Englebretson, R., & Martey, N. ([s.d.]). Santa Barbara corpus of spoken American English, Parts 1-4. *Philadelphia: Linguistic Data Consortium*, 2000–2005.
- Brinton, L. J. (2010). *Grammaticalization and Discourse Functions*. De Gruyter Mouton. <https://doi.org/doi:10.1515/9783110907582>
- Cadiot, A., Ducrot, O., Fradin, B., & Nguyen, T. B. (1985). Enfin, marqueur métalinguistique. *Journal of pragmatics*, 9(2–3), 199–239.
- Camacho, A. (2007). *Swipe: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music* [PhD Thesis]. University of Florida.
- Camacho, A., & Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3), 1638–1652. <https://doi.org/10.1121/1.2951592>
- Campbell, W. N., & Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, 19(1), 37–47. [https://doi.org/10.1016/S0095-4470\(19\)30315-8](https://doi.org/10.1016/S0095-4470(19)30315-8)
- Cavalcante, F. (2020). *The information unit of topic: A crosslinguistic, statistical study based on spontaneous speech corpora* [Tese de doutorado, Universidade Federal de Minas Gerais]. <http://hdl.handle.net/1843/33673>

- Cavalcante, F., & Ramos, A. (2016). The American English spontaneous speech minicorpus: Architecture and comparability. *CHIMERA: Romance Corpora and Linguistic Studies*, 3, 99–124.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. <https://doi.org/10.1037/h0026256>
- Crawley, M. J. (2012). *The R book*. John Wiley & Sons.
- Cresti, E. (2000). *Corpus di Italiano parlato*. Accademia della Crusca.
- Cresti, E. (2010a). La stanza: Un'unità di costruzione testuale del parlato. *Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana*, 713–732.
- Cresti, E. (2010b). La stanza: Un'unità di costruzione testuale del parlato. *Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana*, 713–732.
- Cresti, E. (2020). The pragmatic analysis of speech and its illocutionary classification according to the Language into Act Theory. *In Search of Basic Units of Spoken Language: A corpus-driven approach*, 94, 181–219.
- CRESTI, E., & FUJIMURA, I. (2018). The information structure of spontaneous spoken Japanese and Italian in comparison: A pilot study. *Le lingue extra-europee e l'italiano: aspetti didattico-acquisizionali e sociolinguistici*, 187.
- Cresti, E., Gregori, L., Moneglia, M., MARTINEZ, N., CARLOTTA, M., & Panunzi, A. (2022). *The LABLITA Speech Resources*. 85–108.
- Cresti, E., & Moneglia, M. (1997). L'intonazione e i criteri di trascrizione del parlato adulto e infantile. Em B. Macwhinney (Org.), *Il progetto CHILDES: strumenti per l'analisi del linguaggio parlato*, , Pisa: Edizioni del Cerro, vol (p. 57–90). II.
- Cresti, E., & Moneglia, M. (2010). *INFORMATIONAL PATTERNING THEORY AND THE CORPUS - BASED DESCRIPTION OF SPOKEN LANGUAGE* (Vol. 4).
- Cresti, E., & Moneglia, M. (2018). The illocutionary basis of information structure: The Language into Act Theory (L-Act). Em *Information Structure in Lesser-described Languages* (p. 360–402). John Benjamins.

- Cresti, E., & Moneglia, M. (Eds.). (2005). *C-ORAL-ROM. Integrated reference corpora for spoken Romance languages*. John Benjamins.
- Cuenca, M.-J., & Marín, M.-J. (2012). Discourse markers and modality in spoken Catalan: The case of (és) clar. *Journal of Pragmatics*, 44(15), 2211–2225.
- d’Alessandro, C. (2012). Voice Source Parameters and Prosodic Analysis. Em S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (Orgs.), *Methods in Empirical Prosody Research* (p. 63–88). De Gruyter. <https://doi.org/doi:10.1515/9783110914641.63>
- de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(0), 1917–1930.
- de Melo Rocha, B. N. R. (2016). *Uma metodologia empírica para a identificação e descrição de ilocuções e a sua aplicação para o estudo da Ordem em PB e Italiano*.
- de Ruiter, L. E. (2015). Information status marking in spontaneous vs. Read speech in story-telling tasks – Evidence from intonation analysis using GToBI. *Journal of Phonetics*, 48, 29–44. <https://doi.org/10.1016/j.wocn.2014.10.008>
- Dean, D., Kanagasundaram, A., Ghaemmaghami, H., Rahman, M. H., & Sridharan, S. (2015). The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition. Em H. Ney, E. Noth, S. Moller, B. Mobius, & S. Steidl (Orgs.), *Proceedings of the 16th Annual Conference of the International Speech Communication Association, Interspeech 2015* (p. 3456–3460). International Speech Communication Association. <https://eprints.qut.edu.au/85240/>
- Dean, D., Sridharan, S., Vogt, R., & Mason, M. (2010). The QUT-NOISE-TIMIT corpus for evaluation of voice activity detection algorithms. Em K. Hirose, S. Nakamura, & T. Kaboyashi (Orgs.), *Proceedings of the 11th Annual Conference of the International Speech Communication Association* (p. 3110–3113). International Speech Communication Association. <https://eprints.qut.edu.au/38144/>
- Dedaić, M. N. (2005). Ironic denial: Tobaže in Croatian political discourse. *Journal of pragmatics*, 37(5), 667–683.
- Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP — A collaborative voice analysis repository for speech technologies.

- 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 960–964.
<https://doi.org/10.1109/ICASSP.2014.6853739>
- Dér, C. I., & Markó, A. (2010). A pilot study of Hungarian discourse markers. *Language and speech*, 53(2), 135–180.
- Didirková, I., Crible, L., & Simon, A. C. (2019). Impact of prosody on the perception and interpretation of discourse relations: Studies on “et” and “alors” in spoken French. *Discourse Processes*, 56(8), 619–642.
- Drager, K. K. (2015). *Linguistic variation, identity construction and cognition*. Language Science Press.
- Emmertsen, S., & Heinemann, T. (2010). Realization as a device for remedying problems of affiliation in interaction. *Research on Language and Social Interaction*, 43(2), 109–132.
- Émond, C., Ménard, L., & Laforest, M. (2013). Perceived prosodic correlates of smiled speech in spontaneous data. *Interspeech 2013*, 1380–1383.
- Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor. *Proceedings of the 21st ACM International Conference on Multimedia*, 835–838. <https://doi.org/10.1145/2502081.2502224>
- Ferrari, L., Rocha, B., & Raso, T. ([s.d.]). C-ORAL-ESQ. *in preparation*.
- Ferro, M., & Tamburini, F. (2019). Using Deep Neural Networks for Smoothing Pitch Profiles in Connected Speech. *Italian Journal of Computational Linguistics*, 5(2), 33–48. <https://doi.org/10.4000/ijcol.476>
- Firenzuoli, V. (2003). Verso un nuovo approccio allo studio dell’intonazione a partire da corpora di parlato: Esempi di profili intonativi di valore illocutivo dell’italiano. *Verso un nuovo approccio allo studio dell’intonazione a partire da corpora di parlato: esempi di profili intonativi di valore illocutivo dell’italiano*, 1000–1016.
- Firenzuoli, V., & Signorini, S. (2003). L’unità informativa di topic: Correlati intonativi. *Proc. of XIII Giornate del GFS, Pisa ETS*, 177–184.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378–382.
- Frosali, F. (2008). Il lessico degli Ausili Dialogici. Em *Prospettive nello studio del lessico italiano: Atti del IX Congresso SILFI, Firenze, 14-17 giugno 2006*. - (*Proceedings e report* ; 40). Firenze University Press. <https://doi.org/10.1400/96805>

- Furnival, G. M., & Wilson, R. W. (1974). Regressions by Leaps and Bounds. *Technometrics*, 16(4), 499–511. JSTOR. <https://doi.org/10.2307/1267601>
- Gerhard, D. (2003). *Pitch Extraction and Fundamental Frequency: History and Current Techniques*.
- Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.
- Gerratt, B. R., & Kreiman, J. (2001). Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, 29(4), 365–381. <https://doi.org/10.1006/jpho.2001.0149>
- Gobbo, O. (2019). *Marcadores discursivos em uma perspectiva informal: Análise prosódica e estatística* [Dissertação de Mestrado, Universidade Federal de Minas Gerais]. <http://hdl.handle.net/1843/LETR-BAVN6N>
- Gonzalez, S., & Brookes, M. (2014). PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2), 518–530. <https://doi.org/10.1109/TASLP.2013.2295918>
- Gries, S. T. (2021). *A Practical Introduction*. De Gruyter Mouton. <https://doi.org/doi:10.1515/9783110718256>
- Hakulinen, A. (1998). The use of Finnish nyt as a discourse particle. *Pragmatics and Beyond New Series*, 83–96.
- Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Hastie, T., Mazumder, R., Lee, J., & Zadeh, R. (2014). *Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares*.
- Hermes, D. J. (1988). Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83, 257–264. <https://doi.org/10.1121/1.396427>
- Hilmisdóttir, H. (2011). Giving a tone of determination: The interactional functions of nú as a tone particle in Icelandic conversation. *Journal of Pragmatics*, 43(1), 261–287.

- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6), 82–97. <https://doi.org/10.1109/MSP.2012.2205597>
- Hirst, D., & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Trav. Inst. Phonét. Aix*, 15, 71.
- Hockett, C. F. (1958). *A course in modern linguistics*.
- Izre'El, S., Mello, H., Panunzi, A., & Raso, T. (2020). In search of a basic unit of spoken language: Segmenting speech. Em S. Izre'El, H. Mello, A. Panunzi, & T. Raso (Orgs.), *In Search of a Basic Unit of Spoken Language: A Corpus-driven Approach*. John Benjamins.
- Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71, 1–15.
- Jouvet, D., & Laprie, Y. (2017). Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. *2017 25th European Signal Processing Conference (EUSIPCO)*, 1614–1618. <https://doi.org/10.23919/EUSIPCO.2017.8081482>
- Jucker, A. H. (1997). The discourse marker well in the history of English1. *English Language & Linguistics*, 1(1), 91–110.
- Kasi, K. (2002). *Yet Another Algorithm for Pitch Tracking (YAAPT)*.
- Kawahara, H., Cheveigné, A., Banno, H., Takahashi, T., & Irino, T. (2005). *Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT* (p. 540). <https://doi.org/10.21437/Interspeech.2005-335>
- Keevallik, S., & Vint, K. (2012). Influence of changes in the station location and measurement routine on the homogeneity of the temperature, wind speed and precipitation time series. *Estonian Journal of Engineering*, 18(4).
- Kochari, A. R. (2019). Conducting web-based experiments for numerical cognition research. *Journal of cognition*, 2(1).
- Krifka, M., & Musan, R. (2012). *The expression of information structure* (Vol. 5). Walter de Gruyter.

- Kroon, C. (1997). Discourse markers, discourse structure and Functional Grammar. *Discourse and pragmatics in functional grammar*, 17–32.
- Kruse, J. S., & Barbosa, P. A. (2021). Alinha-pb. *Journal of Communication and Information Systems*, 36(1), 192–199.
- Lavechin, M., Gill, M.-P., Bousbib, R., Bredin, H., & Garcia-Perera, L. P. (2020). End-to-End Domain-Adversarial Voice Activity Detection. *Interspeech 2020*, 3685–3689. <https://doi.org/10.21437/Interspeech.2020-2285>
- Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press.
- Lee, C., Kiefer, F., & Editors, M. K. ([s.d.]). *Studies in Natural Language and Linguistic Theory 91 Contrastiveness in Information Structure, Alternatives and Scalar Implicatures*.
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17), 1–5.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *J. Mach. Learn. Res.*, 18(1), 6765–6816.
- Lumley, T., & Miller, A. J. (2004). *leaps: Regression Subset Selection*. <https://api.semanticscholar.org/CorpusID:135223465>
- Machač, P., & Skarnitzl, R. (2009). *Principles of phonetic segmentation*. Epocha.
- MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.
- Maia Rocha, B., & Raso, T. (2011). A unidade informacional de Introdutor Locutivo no português do Brasil: Uma primeira descrição baseada em corpus. *Domínios de Lingu@gem*, 5(1), 327–343. <https://doi.org/10.14393/DL9-v5n1a2011-21>
- Mallows, C. L. (2000). Some comments on Cp. *Technometrics*, 42(1), 87–94.
- Martin, P. (2015). *WinPitch*. www.winpitch.com
- Martínez, C. N., & Somacarrera, M. L. (2018). Mini-Corpus del español para DB-IPIC. *CHIMERA: Romance Corpora and Linguistic Studies*, 5(2).
- Martínez, N., Somacarrera, C. ; L., & Mini-Corpus, M. (2018). Del espa\textbackslash nol para DB-IPIC. *CHIMERA: Romance Corpora and Linguistic Studies v.*, 5, 197–215.

- Maschler, Y. (1997). Discourse markers at frame shifts in Israeli Hebrew talk-in-interaction. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 7(2), 183–211.
- Mauch, M., & Dixon, S. (2014). PYIN: A fundamental frequency estimator using probabilistic threshold distributions. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 659–663. <https://doi.org/10.1109/ICASSP.2014.6853678>
- Mazeland, H., & Huiskes, M. (2001). Dutch ‘but’ as a sequential conjunction. *Studies in interactional linguistics*, 10, 141–169.
- McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. *Proceedings of the 14th python in science conference*, 8.
- Meer, P., & Fuchs, R. (2022). The Trini Sing-Song: Sociophonetic variation in Trinidadian English prosody and differences to other varieties. *Language & Speech*, 65(4), 923–957. <https://doi.org/10.1177/0023830921998404>
- Mello, H., & Raso, T. (2011). Illocution, Modality, Attitude: Different Names for Different Categories. Em *Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation*. Firenze University Press. <https://doi.org/10.1400/178850>
- Mello, H., Raso, T., Mittmann, M., & Côrtes, P. (2012). Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. Em T. Raso & H. Mello (Orgs.), *C-ORAL–Brasil I: Corpus de referência do português brasileiro falado informal* (p. 125–176). UFMG.
- Mixdorff, H., & Niebuhr, O. (2013). The influence of F0 contour continuity on prominence perception. *Proc. Interspeech 2013*, 230–234. <https://doi.org/10.21437/Interspeech.2013-73>
- Moneglia, M. (2011). Spoken corpora and pragmatics. *Revista Brasileira de Linguística Aplicada*, 11, 479–519.
- Moneglia, M., & Raso, T. (2014). Notes on the Language into Act Theory. Em T. Raso & H. Mello (Orgs.), *Spoken corpora and linguistics studies* (p. 468–494). John Benjamins.
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5–6), 453–467.

- Panunzi, A., & Gregori, L. (2012). An XML database for the representation of information structure in spoken language. Em H. Mello, A. Panunzi, & T. Raso, *Pragmatics and prosody: Illocution, modality, attitude, information patterning and speech annotation* (p. 133–150). Firenze University Press.
- Panunzi, A., & Mittmann, M. (2014). The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. Em T. ; M. and (Org.), *RASO* (p. 129–151). Benjamins p.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct), 2825–2830.
- Plante, F., Meyer, G. F., & Ainsworth, W. A. (1995). A pitch extraction reference database. *Proc. 4th European Conference on Speech Communication and Technology (Eurospeech 1995)*, 837–840. <https://doi.org/10.21437/Eurospeech.1995-191>
- Pradeep, R., Reddy, M. K., & Rao, K. S. (2019). LSTM-Based Robust Voicing Decision Applied to DNN-Based Speech Synthesis. *Automatic Control and Computer Sciences*, 53(4), 328–332. <https://doi.org/10.3103/S0146411619040096>
- Raso, T. (2014). Prosodic constraints for discourse markers. Em T. Raso & H. Mello (Orgs.), *Spoken Corpora and Linguistic Studies* (p. 411–467). John Benjamins. <https://doi.org/10.1075/scl.61.14ras>
- Raso, T., & Cavalcante, F. (2021). The topic information unit: Modeling prosodic forms in a crosslinguistic perspective. *Proceedings. GSCP*.
- Raso, T., & Cavalcante, F. . (2022). Prosódia e estrutura informacional. Em *Oliveira Jr* (p. 40–54). Contexto v. 1, p.. [ACESSO AO LIVRO.
- Raso, T., Cavalcante, F., & Mittmann, M. (2016, junho). Prosodic forms of the Topic information unit in a cross-linguistic perspective: A first survey. *Proceedings. SLI-GSCP International Conference, Rome*.
- Raso, T., & Ferrari, L. (2020). Uso dei Segnali Discorsivi in corpora di parlato spontaneo italiano e brasiliano. Em M. FERRONI R. ;. BIRELLO (Org.), *La competenza discorsiva a lezione di lingua straniera*. Aracne.
- Raso, T., & Mello, H. (2009). 2) Parâmetros de compilação de um corpus oral: O caso do C-ORAL-BRASIL. *Veredas-Revista de Estudos Linguísticos*, 13(2).

- Raso, T., & Mello, H. (2012). *C-ORAL-BRASIL: corpus de referência do português brasileiro falado informal: Vol. I*. UFMG.
- Raso, T., & Mello, H. (2014). *C-oral-brasil: Description, Methodology and Theoretical Framework* (T. BERBER SARDINHA & T. (eds.) SÃO BENTO FERREIRA, Orgs.; Working with Portuguese Corpora). Bloomsbury.
- Raso, T., Rilliard, A., & Santos, S. (2022). Modeling the prosodic forms of Discourse Markers. *Domínios de Linguagem*, 16(4), 1436–1488. <https://doi.org/10.14393/DL52-v16n4a2022-8>
- Raso, T., & Rocha, B. (2016). Illocution and attitude: On the complex interaction between prosody and pragmatic parameters. *Journal of Speech Sciences*, 5, 5–27.
- Raso, T., & Santos, S. (2020). Short information units: A corpus-based prosodic study on the lexeme “assim” in Brazilian Portuguese. *Journal of Speech Sciences*, 8(2), 03–35. <https://doi.org/10.20396/joss.v8i2.14994>
- Raso, T., Teixeira, B., & Barbosa, P. (2020). Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. *Journal of Speech Sciences*, 9(00), 105–128. <https://doi.org/10.20396/joss.v9i00.14957>
- Raso, T., & Vieira, M. A. (2016). Description of Dialogic Units/Discourse Markers in Spontaneous Speech Corpora Based on Phonetic Parameters. *CHIMERA: Romance Corpora and Linguistic Studies*, 3(2), 221–249.
- Reimers, N., & Gurevych, I. (2020, novembro). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. <https://arxiv.org/abs/2004.09813>
- Rocha, B. (2016). *A forma prosódica da locução de Ordem em Português Brasileiro e Italiano: Proposta de metodologia baseada em corpus*. Faculdade de Letras da Universidade Federal de Minas Gerais.
- Rocha, B. M. (2011). *A unidade informacional de introdutor locutivo no português brasileiro: Uma análise baseada em corpus*.
- Rocha, B., Mello, H., & Raso, T. (2019). Para a compilação do C-ORAL-ANGOLA: um corpus de fala espontânea informal do português angolano. *Filologia e Linguística Portuguesa*, 20, 139–157.
- Rocha, B., & Raso, T. (2016). The interaction between illocution and attitude and its consequences for the empirical study of illocutions. *Parler les*

langues romanes/Parlare le lingue romanze/Hablar las lenguas romances/Falando línguas românicas. Napoli: Università di Napoli L'Orientale, 69–87.

- Rofiq, Z. (2018). The study of the Indonesian pragmatic particle *sih*. *LINGUA: Jurnal Ilmu Bahasa dan Sastra, 13*(2), 151–156.
- Rubinsteyn, A., & Feldman, S. (2016). *fancyimpute: An Imputation Library for Python* (0.7.0) [Software]. <https://github.com/iskandr/fancyimpute>
- Santos, S., & Raso, T. (2022). *Hipersegmentação prosódica e processos de morfologização*. II Congresso de Prosódia Brasileiro.
- Sasaki, K., & Yamada, Y. (2019). Crowdsourcing visual perception experiments: A case of contrast threshold. *PeerJ, 7*, e8339.
- Schiffrin, D. (1987). *Discourse markers* (Número 5). Cambridge University Press.
- Shochi, T., Guerry, M., Rilliard, A., Erickson, D., & Rouas, J.-L. (2020). The combined Perception of Socio-affective Prosody: Cultural Differences in Pattern Matching. *Journal of the Phonetic Society of Japan, 24*, 84–96. https://doi.org/10.24467/onseikenkyu.24.0_84
- Steeneken, H. J. M., & Varga, A. (1993). Assessment for automatic speech recognition: I. Comparison of assessment methods. *Speech Communication, 12*(3), 241–246. [https://doi.org/10.1016/0167-6393\(93\)90094-2](https://doi.org/10.1016/0167-6393(93)90094-2)
- Stewart, R., & Sandler, M. B. (2010). Database of omnidirectional and B-format room impulse responses. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 165–168.
- Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior research methods, 42*, 1096–1104.
- Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*(1), 24–31.
- Sturm, N. (2011). *Analyse de la qualité vocale appliquée à la parole expressive* [PhD Thesis]. <http://www.theses.fr/2011PA112021/document>
- Sukhostat, L., & Imamverdiyev, Y. (2015). A Comparative Analysis of Pitch Detection Methods Under the Influence of Different Noise Conditions. *Journal of Voice, 29*(4), 410–417. <https://doi.org/10.1016/j.jvoice.2014.09.016>

- Talkin, D. (2005). *A Robust Algorithm for Pitch Tracking (RAPT)*.
- Talkin, D., & Kleijn, W. B. (1995). A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, 495, 518.
- Talkin, D., Kleijn, W. B., & Paliwal, K. K. (1995). *Speech Coding and Synthesis*.
- Teixeira, B. H. F., & Malvessi Mittmann, M. (2018). Acoustic Models for the Automatic Identification of Prosodic Boundaries in Spontaneous Speech / Modelos acústicos para a identificação automática de fronteiras prosódicas na fala espontânea. *Revista De Estudos Da Linguagem*, 26(4), 1455–1455. <https://doi.org/10.17851/2237-2083.26.4.1455-1488>
- Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2015). Package ‘rpart’. Available online: cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf (accessed on 20 April 2016).
- Traugott, E. C. (1995). The role of the development of discourse markers in a theory of grammaticalization. *ichl xii, Manchester*, 123.
- Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, 107(6), 3438–3451. <https://doi.org/10.1121/1.429414>
- Tucci, I. (2004a). L’inciso: Caratteristiche morfosintattiche e intonative in un corpus di riferimento. *Atti del Convegno Nazionale*, 1–14.
- Tucci, I. (2004b). L’inciso: Caratteristiche morfosintattiche e intonative in un corpus di riferimento. *Atti del Convegno Nazionale*, 1–14.
- Tucci, I. (2009). Obiter dictum: La funzione informativa delle unità parentetiche. *Atti del GSCP*, 3, 635–654.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. Em *Methods in empirical prosody research* (p. 1–28). Mouton de Gruyter.
- Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3), 247–251. [https://doi.org/10.1016/0167-6393\(93\)90095-3](https://doi.org/10.1016/0167-6393(93)90095-3)
- Vaysse, R., Astésano, C., & Farinas, J. (2022). Performance analysis of various fundamental frequency estimation algorithms in the context of

- pathological speech. *The Journal of the Acoustical Society of America*, 152(5), 3091–3101. <https://doi.org/10.1121/10.0015143>
- Venables, W. N., & Ripley, B. D. (2010). *Modern Applied Statistics with S*. Springer Publishing Company, Incorporated.
- Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *J. Phon.*, 48, 1–12.
- Weenink, D. (2022). *Speech Signal Processing with Praat*. <https://www.fon.hum.uva.nl/david/sspbook/sspbook.pdf>
- Xu, Y. (2013). ProsodyPro—A Tool for Large-scale Systematic Prosody Analysis. Em B. Bigi & D. Hirst (Orgs.), *Tools and Resources for the Analysis of Speech Prosody* (Vol. 1, p. 7–10). http://www2.lpl-aix.fr/~trasp/Proceedings/TRASP2013_proceedings.pdf
- Yang, N., Ba, H., Cai, W., Demirkol, I., & Heinzelman, W. (2014). BaNa: A Noise Resilient Fundamental Frequency Detection Algorithm for Speech and Music. *IEEE Transactions on Audio Speech and Language Processing*, 22, 2329–9290. <https://doi.org/10.1109/TASLP.2014.2352453>
- Zahorian, S. A., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123(6), 4559–4571. <https://doi.org/10.1121/1.2916590>

11 APPENDIX A - TRANSCRIPTION CRITERIA

In the following subsections, we present the corpus transcription criteria. This section is based, as a whole, on Mello and Raso (2009) and Mello et al. (2012). We start with some rules applicable to text transcribed in accordance with the standard spelling form [§A.1]. Although this is to some extent self-explaining, it needed some standardization. We then pass to non-linguistic criteria ([§A.2] through [§A.10]). From [§A.11] to [§A.13], introduce the criteria for linguistically less conventionalized phenomena. From [§A.14] on, we present the criteria concerned with the core linguistic phenomena, following, when possible, the order *phonetics/phonology* > *morphology* > *syntax* > *lexicon*.

Phenomena are provided with examples from the corpora and respective translations. We tried to keep translations as close as possible to originals, from a structural standpoint, to facilitate the comprehension of phenomena. Thus, translations sometimes should not be taken literally. The audio files presented in this work are available at <[SHARED MATERIALS THESIS](#)>.

Before delving into them, we want to call the reader's attention to an issue. Each section represents a criterion according to which transcription errors were tallied. Some criteria encompass subcriteria. In a perfect scenario, the sections would have been broken down until each phenomenon was completely homogeneous. However, this would lead to a much more complex work, which we decided not to undertake at this moment.

1. Standard spelling form

The Brazilian Portuguese (BP) standard norm was observed in all cases where a special criterion lacked. Words are transcribed following the

BP standard spelling form as pronounced, i.e., without the inclusion or exclusion of items. The standard spelling form followed the Orthographic Form of 1943. The Houaiss Dictionary of the Portuguese Language (Houaiss, 2nd Edition, April 2007) was chosen to be the reference for spelling forms in as much as it was the most complete Portuguese dictionary to that date. The following mistakes were tallied in this criterion.

1.1 Misspelling

It accounts for general misspelled form of all tokens for which there is no special criterion.

1.2 Words unintentionally spelled in accordance with the new spelling form (Orthographic Agreement of 1990), like:

- a) *lingüiça* (sausage) instead of *linguiça*;
- b) *freqüência* (frequency) instead of *frequência*;
- c) *jibóia* (boa) instead of *jiboia*;
- d) *mini-projeto* (mini project) instead of *miniprojeto*;
- e) *mão-de-obra* (manpower) instead of *mão de obra*;
- f) *microondas* (microwave) instead of *micro-ondas*;

1.3 Capitalization or syllable division

Except for titles and proper nouns, words should never be capitalized, even at the beginning of the terminated sequence. Syllable division is not used unless the word is scanned and separated by a prosodic boundary sign. Such occurrences are commented on the metadata.

1.4 Pronunciation mistakes

If the speaker incorrectly pronounces a word and correct herself

subsequently, the mistake is transcribed as pronounced. If the speaker, on the other hand, mistakes the pronunciation without repairing it, the standard spelling form is transcribed and commented on the metadata.

1.5 Alphabet letter names

Alphabet letters are transcribed orthographically. If, say, *letter j* is referred to in the audio, it is transcribed as *letra jota* (letter jay).

2. Word misunderstanding

Misunderstood words, word deletion, and word insertion are set off from the main standard spelling form criterion and tallied separately.

Word misunderstanding [bmedrp01_1_057]

*GIU: [57] a gente quer / por exemplo / o lance do devedê **do**
Metropolitan //

*GIU: [57] we want / for instance / the DVD thing **of the**
Metropolitan //

In the example below, the transcription displays *do* (of the), whereas what is actually pronounced is *no* (at the).

3. Word deletion

Word deletion accounts for words that, although present in the audios, are not transcribed. The example below presents a transcription where the word *ai* (then) is not transcribed.

Word deletion [btelpv31_099]

*JES: [99] <&a [/1] aí eu> tava pertinho / **(aí)** eu aproveitei e fui lá //

*JES: [99] <&th [/1] then I> was around / **(then)** I took the occasion and went there //

Any deleted word, notwithstanding the concurrent applicability of other special criteria, is counted in this type of error.

4. Word insertion

Word insertion comprises words transcribed despite not being present in the audio, such as *eu* (I) in the example below.

Word insertion [bnatpd10_019]

*ARN: [19] <não> / eu tô [/2] eu nũ tô vendo não / mas **eu** tô percebendo //

*ARN: [19] <no> / I'm [/2] I can't see it / but **I** can notice //

5. Unintelligible speech and anonymization

Words and speech chunks heard but not understood by transcribers receives a special sign. If just one word was not identified, it is transcribed with the symbol xxx.

Unintelligible word [bnatpd07_068]

*COA: [68] essa tréplica dela acaba **xxx** //

*COA: [68] her rejoinder ends up **xxx** //

When more than one word is not identified, the speech chunk is transcribed with *yyyy* (second example below).

Unintelligible speech chunk [bnatbu03_392]

*DBC: [399] *cê entendeu / **yyyy** / ah / tem que ficar não sei o que parado na conta / pode aplicar //*

DBC: [399] did you get it / **yyyy / say / anything must remain in the account for some time / you can invest it //*

In some cases, the audio is beeped so as to preserve participants' anonymity and/or privacy. Anonymized speech chunks are transcribed with the symbol *yyy*. The example below displays the anonymized number of a legal proceeding. The fact that the participant lives with HIV is the main reason why the proceeding was brought by.

Anonymization [bnatpd07_068]

*ANI: [1] audiência de instrução e julgamento / do processo número **yyy** / **yyy** / **yyy** / **yyy** / **yyy** //

ANI: [1] evidentiary and judgment hearing / case number **yyy / **yyy** / **yyy** / **yyy** //*

6. Paralinguistic and non-linguistic elements

Cough, gasps, groans, laughs, moans, sighs, throat-clear, as well as other non-linguistic sounds produced or referred to by participants are transcribed with the symbol *hhh*.

Paralinguistic sound [bteplv22_050]

*MOI: [50] **hhh** cê é doido **hhh** //

*MOI: [50] **hhh** you're mad **hhh** //

Paralinguistic and non-linguistic elements are not set off from the ongoing speech stream with prosodic boundary signs unless they are really prosodically parsed and bear communicative value at the same time, such as non-natural, ostensive coughs, laughs and even surprise or astonishment sounds. The example below has a paralinguistic sound that emulates surprise used as an answer to a rather banal fact.

Paralinguistic sound prosodically parsed [bteplv27_010-012]

*GRA: [10] ah / foi bom / mãe / cê acredita que só fui eu e &um
[/1] mais uma menina //

*LUZ: [11] **hhh** // [12] por quê //

*GRA: [10] oh / it was good / mom / do you believe that only me
and &an [/1] another girl showed up //

*LUZ: [11] **hhh** // [12] why //

Due to their high frequency and degree of conventionalization, two paralinguistic sounds have their own sign. The first one is the dental click sound used to mark annoyance, which is transcribed with *nts*:

Click sound [bnatla05_230]

*AGN: [231] como é que é / **nts** / &he / &sa [/1] Sagrada Família
//

*AGN: [231] *how is that* / **nts** / &he / &sa [/1] *Sagrada Família*
//

Likewise, the whistle-like sound generally used to get someone's attention or to shush someone is transcribed with *psiu* (loosely translatable as *psst* or *shh*).

Psiu sound [bmedts01_1_266]

*MRC: [264] *vai fazer aquilo de verdade* // [265] &el [/1] *ele nũ*
+ [266] **psiu** // [267] *tũ brincando hhh* //

*MRC: [264] *you're gonna do that for real* // [265] &he [/1] *he*
doesn't + [266] **psiu** // [267] *I'm kidding hhh* //

7. Retraction

As explained in [\[§2.3\]](#), retractions are frequently accompanied by non-terminal prosodic boundaries. When this is the case, the boundary sign receives the additional annotation of how many words are retracted by the speaker. One additional aspect of this annotation that needs to be heeded is that the number of words canceled out must be distributed among boundary signs whenever the retraction scope goes back beyond one prosodic unit (speech chunk between two boundaries):

Wrong annotation of retraction [bmedts07_175]

*JOS: [175] *mas o que &diz / já que você <nũ>* **[/8]** *<o que que*
cê tava> <falando> //

*JOS: [175] *but what did you &say / since you're <not>* **[/8]**
<what were you> <talking about> //

Correct annotation of retraction [bmedts07_175]

*JOS: [175] mas o que &diz [/**4**] já que você <nũ> [/**4**] <o que que cê tava> <falando> //

*JOS: [175] *but what did you &say [/**4**] since you're <not> [/**4**] <what were you> <talking about> //*

Since validating the corpus segmentation was out of the scope of this work, just the number of retracted words and their distribution were validated.

8. Numerals

Numerals are transcribed in accordance with the standard spelling form, and the Arabic numeral symbols (numerical digits) are left for the mark-up and annotation scheme. The transcription followed some particularities.

8.1 Hyphenation

For the sake of comparability with other C-ORAL corpora, numerals, either cardinals or ordinals, are hyphenated and counted as a unique word.

Ordinals [bmedsp02_305]

*AND: [305] a distância do Vasco pro **décimo-sexto** / era de quatro / e nũ é / de seis / né //

*AND: [305] *Vasco's distance to the **sixteenth** classified / was four / and not / six / huh //*

8.2 Non-hyphenation of approximators/estimators

Non-numeric expressions used to approximate or estimate a number, such as *e pouco* (something) or *e sei lá* (and whatever), are not hyphenated together.

Approximation/estimation of numbers [bnatbu03_392]

*DBC: [392] é cento-e-quarenta **e poucos** / cento-e-trinta **e poucos** / *nũ sei //*

*DBC: [392] *it's one-hundred-forty something / one-hundred-thirty something / I don't know //*

8.3 Non-hyphenation of decimals and thousands separators

When they are pronounced, decimal and thousands separators are not hyphenated together with numerals.

Separators [bnatpd01_014]

*NEW: [14] só no mês de março / **quarenta-e-um vírgula nove //**

*NEW: [14] *only on march / forty-one comma nine //*

8.4 Hyphenation of fractional numerals

Only the part forming a regular cardinal numeral is hyphenated. The table below displays some examples:

Examples of fractional numerals

Numeral	Portuguese standard norm	Corpus convention	Translation
$1 \frac{1}{2}$	Um e meio	um e meio	One and a half
$21 \frac{1}{2}$	Vinte e um e meio	vinte-e-um e meio	Twenty-two and a half
$\frac{3}{4}$	Três quartos	Três quartos	Three thirds
$21 \frac{1}{40}$	Vinte e um quadragésimos	Vinte-e-um quadragésimos	Twenty-one fortieths
$21 \frac{1}{41}$	Vinte e um quarenta e um avos	Vinte-e-um quarenta-e-um avos	Twenty-one forty-firsts

The fractions of hours follow, thus, this rule.

Fractional numerals [btelpb13_005]

*SAN: [5] bom / então / olha só // [6] minha cliente das cinco e meia acabou de desmarcar / cê quer vim **cinco e meia** //

*SAN: [5] well / so / look // [6] my half-past-five costumer just called to cancel / do you wanna come **at half past five** //

8.5 Numerals spelled digit by digit

Numerical codes or numerals recombined for simplification were transcribed as pronounced. If digit by digit, no hyphenation is used. If a recombination, just the recombined numerals are hyphenated.

Numerals spelled digit by digit [btelpv38_032-033]

*ALE: [32] pode falar / Marcelo //

*MAR: [33] **nove nove dois** /

*ALE: [32] *say it / Marcelo //*

*MAR: [33] **nine nine two** /

8.6 Numerals separated by prosodic boundaries

When one numeral is separated by a prosodic boundary, just the separated parts are hyphenated.

Compound number divided into two prosodic units
[bnatbu04_036]

*GER: [36] **vinte-mil** / <**e-quatrocentos** / o> valor //

*GER: [36] **twenty-thousand** / <**four-hundred** / the> value //

9. Hesitations and interrupted words

9.1 Hesitations

Hesitation are transcribed as *&he* no matter the vowel quality the sound is produced with. They are transcribed as many times as they are effectively produced and separated by a boundary sign just in case they are prosodically parsed.

Hesitation [bmedts06_026]

*CAR: [26] **&he** / **&he** / a pesquisa pega / meninos / e meninas
/ <né> //

*CAR: [26] **&he** / **&he** / the research considers / boys / and girls

/ <huh> //

9.2 Cases of hesitations that needed to be heeded

Transcribers received instruction to pay close attention to the forms *eh* (interjection) and *é* (it is or yes), whose sounds can be confounded with that of hesitations. To decide which form should be employed, transcribers could both use prosodic cues and check the adequacy of replacing the symbol by yes (since *é* is frequently used to convey agreement) or by a different interjection, as a commutation test. It was agreed that, whenever the sound could convey agreement, the form *é* would take preference.

Hesitation [btelpb16_007-008]

*KEN: [7] deixa eu te falar // [8] **&he** / a Cléo já voltou //

*CLA: [9] já sim //

*KEN: [7] let me ask you // [8] **&he** / did Cléo come back //

*CLA: [9] she did //

Agreement [bmedts01_02_069-071]

*SUE: [69] <isso / é verdade> hhh //

*FAT: [70] **é** //

*SUE: [69] <is that / true> hhh //

*FAT: [70] **yes** //

Interjection [bmedts04_278]

*ANG: [278] <eh> / que coisa <boa> //

*ANG: [278] <ahh> / what a good thing //

9.3 Interrupted words

Interrupted words are also signaled by the ampersand symbol & so as to enable their identification and exclusion from word counts. Although sometimes the full word can be identified with the help of context, only the pronounced part is transcribed.

Interrupted word [bmedts06_026]

*FRE: [117] <tá> // [118] cê acha que cê **volta** / antes das oito / lá do [/1] do [/1] do Del Rey //

*MAR: [119] **&vol** [/1] claro / meu filho //

*FRE: [117] <okay> // [118] you think you **come back** / before eight / from [/1] from [/1] from the Del Rey //

*MAR: [119] **&vol** (I come, as a resumptive answer conveying agreement) [/1] sure / son //

10. Acronyms

In BP, acronyms are usually pronounced in two fashions. When the letters forming the acronym match a possible syllabic combination of the language, the acronym is pronounced as a regular word. Otherwise, each letter is spelled out to form a word. The transcription of acronyms followed, thus, this twofold criterion and has some peculiarities.

10.1 Word acronyms

Acronyms of the former type were simply transcribed in uppercase letters as they are usually written. In the example below, CEMIG stands for the *Companhia Energética de Minas Gerais* (Energy Company of Minas Gerais).

Regular acronym [bnatla01_164]

*CLA: [164] vocês prestaram pra **CEMIG** / e pra Receita Federal //

*CLA: [164] you guys worked for **CEMIG** / and for the Federal Revenue //

10.2 Acronyms spelled letter by letter

Acronyms of the latter type received a special convention. Each letter should be transcribed orthographically as it is pronounced and put together in lowercase. Most Portuguese letter names receive an accent mark in written language, such as *pê* for *p*. Since they are also spelled as a unique word, acronyms of this type respect the accentuation rules of regular Portuguese words. This system, which is out of the scope of this work, provides rules mainly for proparoxytone and oxytone words, the latter encompassing most acronyms of this type. For instance, the correct transcription of *PDV* (which stands for *Plano de Demissão Voluntária*, in English, Volunteer Dismissal Program) is *pedevê* because the accent falls on the last syllable, and not *pêdêvê*.

Spelled-out acronym [bnatla01_184]

*EDU: [186] que a &P [/2] a / Protex diz que tava usando o **ceenepejota** da Confederal //

*EDU: [186] that &P [/2] (the) / Protex claims they were using the **ceenepejota** of Confederal //

In the example above, *cenepejota* stands for *Cadastro Nacional de Pessoas Jurídicas* (National Register of Legal Entities) and it is formed by assembling *cê* (c) + *ene* (n) + *pê* (p) + *jota* (jay). In this case, the accent falls on *jo* and there is no need for an accent mark according to the accentuation rules of Portuguese, whose words are paroxytone by standard.

10.3 Mixed acronyms

Some acronyms may be pronounced in a mixed fashion or be accompanied by numerals. One such case is MPEG-4. The first letter is spelled out (*eme* for *M*) and the remaining part is pronounced like a regular word (*pegue* for *PEG*). In this case, the acronym is transcribed as pronounced, thus *emepegue*, in lowercase and in a unique word. Numerals, in their turn, are transcribed separately following their own criteria [§A.6]. MPEG4 is, thus, transcribed as *emepegue quatro*. If there were, say, an MPEG21, it would be transcribed as *emepegue vinte-e-um*.

Two final observations on this criterion. The first one is that acronyms that have become full words – taking on inflections and being regularly written with lowercase letters – are transcribed as regular words. Thus, *radar* (radar) is not uppercased and *óvni* (UFO) receives an accent mark. Finally, *OK*, which is spelled out in accordance with the original English letter names, is adapted as *oquei*.

11. Foreign words

Foreign words and foreign proper nouns are transcribed in accordance with their original spelling forms.

Foreign word [bmedrp09_2_007]

*JUL: [7] a primeira barreira foi tecnológica // [8] muitos taxistas não sabiam nem atender uma ligação num **smartphone** //

*JUL: [7] *the first was a technological barrier* // [8] *many taxi drivers could barely answer a call using a **smartphone*** //

Minor phonological adaptations on the pronunciation of foreign words, like the paragoge of [i] at consonantal syllable codas, are not transcribed unless the detail is referred to by speakers. On the other hand, if the word is clearly pronounced incorrectly, it is transcribed with the wrong spelling. In such cases, transcribers must provide this information on the metadata.

Foreign word incorrectly pronounced [bfamdI04_047]

*SIL: [47] como se a gente tivesse num **Big Brogher** //

*SIL: [47] as if we were on **Big Brogher** //

Loanwords that have already undergone phonological and orthographic adaptations to enter the Portuguese lexicon, such as *clip*>*clipe*, *stress*>*estresse*, and *portfolio/portafoglio*>*portfólio*, are transcribed in accordance with the Portuguese spelling form unless they are pronounced as in their original languages. In this case, the original spelling form is used.

12. Onomatopoeias

Many onomatopoeias are already conventionalized, either formally or by the use, and the transcription tended to follow these conventions. They follow, anyway, the pronunciation. For instance, the onomatopoeia of knocking a door can be transcribed either as *toc toc* or *toque toque*, the latter preferred when there is the paragoge of [i]

after [k], the most frequent case.

Onomatopoeia [bmedsp02_132-135]

*PAU: [132] eu esqueci // [133] como é que é //

*AND: [134] **pum / pum** //

*PAU: [135] ah // [136] isso //

*PAU: [132] *I forgot* // [133] *how it is* //

*AND: [134] **poom / poom** //

*PAU: [135] *oh* // [136] *that's it* //

In some situations, it may happen that one of the participants is reading out a text for the others. Readers may sometimes use the sound *nanananã* so as to signal that some part of the text is skipped for not being of interest. It is transcribed like this no matter how many times the syllable *na* is repeated.

Skipped text sound [bpucv02_154]

*OSV: [154] em vistoria realizada no dia quatorze do sete **nanananã** / no endereço acima mencionado / constatamos uma residência / que dista +

*OSV: [154] *in an inspection carried out on the fourteenth day of the seventh* **nanananã** / *at the above-mentioned address / we found a residence / that is +*

13. Interjections and exclamations

The interjections *ah*, *eh*, *ih*, *oh*, *uh* and exclamations *oi* (hi), *olá* (hello), *alô* (hello) are transcribed in accordance with the standard spelling form. The vocative exclamation frequently used in PB is transcribed as *ô*, which is similar to the old, poetic English vocative form *O* but rather frequent in BP spontaneous speech. The distinction between this and other forms may present some difficulty, which is addressed in [\[§A.13.1\]](#). Likewise, a set of aspirated and glottalized sounds used for multiple purposes received special conventions [\[§A.13.2\]](#) and require attention. Exclamations of religious genesis also received special transcription rules [\[§A.13.3\]](#). Finally, plural marks were respected when they are pronounced [\[§A.13.4\]](#).

13.1 Distinction between *oh*, *ô*, and *o'*

The vocative exclamation frequently precedes names of persons being addressed. Its quality may vary depending on the diatopy and context but it is always transcribed as *ô*. The distinction between *ô* (vocative), *oh* (interjection), and *o'* (reduced form of the verb to see [\[§A.23\]](#)) may sometimes be difficult. Transcribers were, thus, instructed to replace it by another interjection (such as *ah*) and by the full form of *o'*, *olha* (look), to check which one was more suitable for the context.

Vocative exclamation [btelpv38_016-018]

*MAR: [16] **ô** Alex // [17] <enquanto> eu procuro aqui /

*ALE: [18] <oi> //

*MAR: [16] Alex // [17] <while> I'm searching here /

*ALE: [18] <what> //

13.2 Aspired and glottalized sounds

Some exclamations are often employed to express agreement, disagreement, irony, doubt, as well as to show that the discourse is being followed and understood. These exclamations, namely *hum*, *ham*, *uhn*, and *ahn*, are transcribed in accordance with their pronunciation. To decide between them, transcribers should check the consonantal sound (aspired or glottalized) and the vowel quality. Aspired sounds are transcribed either as *hum* or *ham*, and glottalized sounds as *uhn* or *ahn*. The use at context may also help to distinguish between them. *Hum hum* and *ham ham* are frequently used to show agreement or that the hearer is following the discourse. Some examples help clarify the distinctions.

hum hum used to agree [btelvp03_042-044]

*BRU: [41] <entendeu> // [42] <o instrutor> é [/1] é meu conhecido / a gente combinou assim //

*GAB: [43] ah sim // [44] **hum hum** //

*BRU: [41] <you got it> // [42] <the instructor> is [/1] he's an acquaintance of mine / we agreed this way //

*GAB: [43] oh okay // [44] **hum hum** //

On the other hand, *uhn uhn* and *ahn ahn* are frequently used to express disagreement.

uhn uhn used to disagree [btelpb04_029-030]

*RON: [29] então nũ vai ser a quantidade de vias mais um não

//

*JON: [30] **uhn uhn** //

*RON: [29] *so it won't be as many copies as [promissory] notes plus one //*

*JON: [30] **uhn uhn** //

Isolated, *hum*, *ham*, *uhn* and *ahn* may be used to express doubt, comprehension, irony, and to show that the hearer is following the discourse or some instructions, depending on the prosodic realization.

Ahn used to express doubt [btelpv05_024-027]

*REN: [24] *cê almoçou &f +*

*TER: [25] **ahn** //

*REN: [26] *cê foi almoçar fora //*

*TER: [27] *fomos almoçar fora / menino //*

*REN: [24] *you went out to &l +*

*TER: [25] **ahn** //

*REN: [26] *you went out to lunch //*

*TER: [27] *we went out to lunch / girl //*

12.3 Exclamations of religious genesis

Exclamations of religious genesis – like *Nossa Senhora* (Our Lady), *Virgem Maria* (Virgin Mary), *Ave Maria* (Hail Mary), or *Jesus* – are rather

frequent in BP and, thus, received special conventions. Firstly, they are always capitalized. Some full forms often take on reduced forms and should be transcribed as such. *Nossa Senhora* (Our Lady) may become *Nossa*, *No'*, *Nu'*, and even a form redeveloped to reinforce perplexity, *Nusga*. *Virgem Maria* (Virgin Mary) may be transcribed as *Vixe'* or *Vix'*, depending on whether the final vowel is realized. And *Ave Maria* (Hail Mary) may be reduced to just *Ave* or *Aff'*.

12.4 Plural mark

Although invariable according to the traditional grammar, interjections and exclamations may occasionally take on the plural form in speech. Therefore, it was conventionalized that both should receive the plural mark (-s) when it was pronounced. *Olá* (hello) may, thus, become *olás*, *oi* (hi) → *ois*, and *ô* (*o vocative*) → *ôs*.

14. Rhotacism

As aforementioned, phenomena of phonetic-phonologic nature were left out of the transcription criteria. The exception is rhotacism. As Mello et al. (2012) point out, this phenomenon is rather common and perceptually salient in PB, especially in lower diastratic varieties. It may happen at consonantal clusters like /bl/, /kl/, /fl/, /gl/, /pl/, /tl/, /vl/, the /l/ shifting to /R/. It may also happen at the syllable coda, such as, for instance, in *vol.tou* > *vor.tou* (it came back). Although it does not necessarily imply lexicalization or grammaticalization processes, rhotacism is respected in the transcription. Thus, if, say, *atlético* (athletic) is pronounced as *a[tr]ético*, it is transcribed *atrético*.

Rhotacism in stop cluster [bfamcv11_106]

*TIT: [106] esses remédio que eu tenho costume de tomar não me **comp^rica** (complica) //

*TIT: [106] these medicines I usually take don't **do** me **harm** //

Rhotacism in syllable coda [bfamnm14_100]

*ANT: [100] aí nós tava chegando aqui no **arto** (alto) aqui descendo //

*ANT: [100] *so we were arriving on the **heights** going down here* //

15. Number agreement in verbs

There is a well-known tendency in BP for subject pronouns to be cliticized and retained by the verb and for plural verb forms to become less used. The inflection used by speakers is respected in the transcription.

Non-standard first-person plural [bnatps11_049]

*PED: [160] <se ele fosse> morrer / **nós** nã **ia** (*nós íamos*) botar o Sarney de vice //

*PED: [160] <if he were going> to die / **we wouldn't** have Sarney as vice [-president] //

Non-standard second-person plural [bnatps11_049]

*CAR: [49] **cês pode** (*cês podem*) ver que estamos ali /

*CAR: [49] **you can** see us there /

Non-standard third-person plural [btelpb04_069]

*JON: [69] **es** <**vai** (*vão*) imprimir> só o recibim com a

promissória embaixo //

*JON: [69] **they** <will print> just the receipt and promissory note below //

Reduced plural forms, like *foro* (they went), are transcribed in accordance with the standard spelling form, in this case, *foram*.

16. Number agreement in nouns and adjectives

Another tendency is for nouns and adjectives to lose the plural morph, which is retained only by the article. The absence of plural morph is also respected in the transcription.

Non-standard noun plural [btelpb04_069]

*ROB: [21] sai mais barato / ajudar **os argentino** (*argentinos*) a resolver o problema do default /

*ROB: [21] *it pays off / helping **the Argentinians** to solve the default problem /*

Non-standard noun and adjective plurals [bmedpr08_2_059]

*ENA: [59] que é **as barca antiga** //

*ENA: [59] *which are **the old boats** //*

17. First-person plural verbal variant forms

The first-person plural verbal inflection may be marked by a reduced

form. The transcription follows the pronunciation. Transcribers should, thus, observe two aspects of the form: the thematic vowel (underlined in the examples below) and the realization of the final /s/. For instance, the form *-amos* may be replaced either by *-amo* or *-emo*. Likewise, *-emos* > *-emo*, *-imos* > *-imo*.

First-person plural verbal variant form [bmedts02_151-152]

*PED: [151] aí nós fizemos as Diretas Já / **ganhamo** (ganhamos) // [152] aí **fizemo** (fizemos) a eleição //

*PED: [151] so we led the Diretas Já / **we won** (ganhamos) // [152] and **we hold** the elections //

18. Variant forms of the verb *estar* (to be)

The verb *estar* (to be) may lose its first syllable *es-* virtually in any form and must be transcribed accordingly.

Variant form of *estar* (to be) [btelpv33_126]

*DON: [126] eles **tavam** (*estavam*) brigando / coitada //

*DON: [126] they **were** having an argument / the poor thing //

Although inflections follow, in general, the standard norm, two forms received minor modifications. Firstly, the apheresis of *estou*³¹ (I am) is transcribed as *tô* (instead of *tou*) so as to follow the use in informal written BP. Secondly, the apheresis of *esteja*³² (be) is transcribed either

³¹ Present tense indicative first-person singular form.

³² Form of the present tense subjunctive first- and third-person singular and of the imperative

as *teje* or *teja*, depending on the pronunciation.

19. Variant forms of the verb *ir* (to go)

The present tense indicative first-person singular standard form of the verb *ir* (to go) is *vamos* (we go). This form is frequently used with a cohortative function, similar to the use of *let's* in English, and is oftentimes reduced to *vamo* or *vãõ*. The reduced forms are respected in the transcription.

Variant forms of verb *ir* (to go) [bmedsp03_159]

*DEN: [159] nós **vãõ** (*vamos*) trocar umas idéia //

*DEN: [159] we're **gonna** bounce some ideas off each other //

The form *vãõ* is also shared with the indicative third-person plural form.

20. Variant forms of the verb *vir* (to come)

Portuguese infinitive verbal forms are marked by a final /r/, which is frequently lost in speech. The infinitive form of the verb *vir* (to come) may additionally be nasalized, coinciding with the form *vim* (I came). This variation is transcribed in accordance with the pronunciation.

Variant forms of verb *vir* (to come) [btelpb28_009]

first-, second- and third-person singular.

*SAN: [9] então pode **vim** (vir) //

*SAN: [9] so you may **come** //

21. Variant forms of the verb *ter* (to have)

The present tense indicative first-person singular form of the verb *ter* (to have), *tenho* (I have), takes on a reduced variant form, *tem*, especially in the phrase *eu tem que* (I have to). This form is shared with the third-person *ela tem que* (she has to).

Variant form of the verb *ter* (to have) [bpubdl02_238]

*JAN: [238] depois **eu tem** que comprar uma //

*JAN: [238] later on **I have to** buy one //

22. Variant forms of the verb *poder* (can)

The present tense indicative first-person singular form of the verb *poder* (can/may), *pode* (I can / I may) frequently takes on a reduced form *po'*, which is transcribed following the pronunciation.

Variant form of the verb *poder* (to have) [bnatpr09_143]

*ANT: [143] **po'** (pode) ficar tranquilo //

*ANT: [143] you **can** rest assured //

23. Variant forms of the verb *olhar* (to look)

The imperative second-person singular form of the verb *olhar* (to look), *olha* (look), also take on two apocopated forms. The form may be

transcribed either as *a'* or *o'* depending on the quality of the vowel pronounced.

Variant form of the verb *olhar* (to look) [bmedin01_2_092]

*JMM: [92] **a'** lá //

*JMM: [92] **look** over there //

Variant form of the verb *olhar* (to look) [bnatla02_077]

*JOS: [77] tá pra cá **o'** (olha) //

*JOS: [77] it's over here **look** //

24. Variant forms of the verb *tomar* (to take)

The imperative second-person singular form of the verb *tomar* (to take), *toma* (take), also has an apocopated form, transcribed as *tó*.

Variant form of the verb *tomar* (to take) [bfamd133_157]

*HER: **tó** / vai guardando / isso aí //

*HER: **take it** / keep putting away / that //

25. Contraction of prepositions and articles

25.1 Standard norm contractions

The Portuguese standard spelling form provides for the contraction of a few prepositions and articles. For instance:

a (to) + articles

ao, à, aos, às

de (of/from) + articles

do, da, dos, das, dum, дума, duns, dumas

em (in/on/at) + articles

no, na, nos, nas, num, numa, nuns, numas

por/per (by/for) + articles

pelo, pela, pelos, pelas

25.2 Special additional contractions

The transcription follows the pronunciation, allowing for contractions not covered by the standard norm. Some frequent non-standard contractions are:

com (with) + articles

co, ca, cos, cas, cum, cuma, cuns, cumas

para (for/to) + articles

pra, pro, pras, pros, prum, pruma, pruns, prumas

The text below brings an example of a non-standard contraction recorded in formal context.

Non-standard contraction of preposition and article
[bmednw06_085]

*CAR: [85] nós estamos vivendo um momento / &he / aonde / né / a / população fala **cos** (*com + os*) parlamentares / os parlamentares trazem ao relator as suas sugestões / e / é natural que nesse momento aconteçam ajustes //

*CAR: [85] *we're living a moment / &he / in which / huh / the / population speak **with (the)** representatives / representatives bring suggestions to rapporteurs / and / it's natural to have some adjustments in these moments //*

25.3 Additional variant forms of prepositions and their contractions

Two prepositions also received variant forms. *Para* (for/to) may be reduced to *pa* or *p'*. Like the others, this form may contract with the articles.

pa/p' (for/to) + articles

po, pos, pa, pas, pum, puns, puma, pumas

The preposition *em* (in/on/at) may also take on the form *ni*. The contractions of this form with the articles results in forms already covered by the standard norm.

Non-standard form of *em* (in/on/at) [bmedex13_51]

*MAR: [51] e aí / o primeiro tempo foi muito ruim / foi / &he / ruim **ni** todos os aspectos / né //

*MAR: [51] so / the first half went pretty badly / it was / &he / bad **in** all aspects / right //

26. Contraction of prepositions and other words

26.1 Standard norm contractions

The BP standard norm also provides for the contraction of some prepositions with other words, like pronouns, demonstratives, and some adverbs. Without being exhaustive, we present the most frequent:

a (to) + aquele/aquela (that)

àquele, àquela, àqueles, àquelas

de (of/from) + ele/ela (he/she)

dele, dela, deles, delas

de (of/from) + aqui (here) / ali (there)

daqui, dali

em (in/on/at) + esse/essa (this)

nesse, nessa, nesses, nessas

em + outro/outra (other)

noutro, noutra, noutros, noutras

26.2 Special contractions

Some contractions not covered by the standard norm are allowed by the corpus transcription rules so as to adapt to the pronunciation. For this, a reduced form of the preposition followed by apostrophe is used. The reduced forms are *c'* (*com*), *d'* (*de*), *n'* (*em*), *p'* (*para*), *pr'* (*para*). They are, all the same, separated by a space from the words they contract with.

Non-standard contraction of preposition and subject pronoun [btelpb29_069]

*BRU: [69] então nã precisa **d' eu** (de eu) preocupar não //

*BRU: [69] so there's no need **for me** to worry //

Non-standard contraction of preposition and demonstrative pronoun [bnatla03_123]

*ALE: [123] você mora **p' aquela** (para aquela) <região> //

*ALE: [123] you live **over that** <region> //

The ways contractions with the above-mentioned prepositions can happen are not provided for beforehand. This is an open-list criterion that allows for as many combinations as found in the corpus.

26.3 Contractions with non-standard variant forms of pronouns and demonstratives

The subsections to follow introduce non-standard variant forms for second- [\[§A.27\]](#) and third-person pronouns [\[§A.28\]](#), as well as for reduced demonstratives [\[§A.29\]](#). The way these forms contract with prepositions depends on which rule ([\[§A.26.1\]](#) or [\[§A.26.2\]](#)) applies to the contraction of their standard variant form.

Suppose the contraction of the preposition *de* (of/from) and the reduced demonstrative variant form *aques* (those) [\[§A.29\]](#). Its standard form is *aqueles* (those). There is, indeed, a contraction provided for by the standard norm, which is *daqueles*. In this case, the contraction of the non-standard form follows [\[§A.26.1\]](#).

Contraction following [\[§A.26.1\]](#) [btelpv29_156]

*SEB: [115] tem um fusquinha verde / na porta da loja aqui / um fusquinha **daques** antiguim mesmo / verdim / original //

*SEB: [115] *there's a green Beetle (car) / in front of the store's doorway / one **of those** very old Beetles / all green / original //*

Now suppose we have the contraction of the preposition *com* (with) and the second-person singular non-standard variant form *ocê* (you). This contraction follows [\[§A.26.2\]](#) since the contraction of *com* + *você* (the standard form) is not covered by the standard norm.

Contraction following [\[§A.26.2\]](#) [btelpb29_039]

*BRU: [39] a Aline conversou **c' ocê** (com ocê) sobre o lanche que vai ter que ter todo dia //

*BRU: [39] *did Aline talk **to you** about the snack supposed to be served the whole period //*

27. Second-person pronoun variant forms

The second-person pronoun *você* (you) can also be transcribed with its reduced forms. They are, namely, *ocê/ocês* and *cê/cês* (you/you all).

Non-standard second-person pronoun variant form
[bnatbu03_275]

*DBC: [275] **cê** entendeu //

*DBC: [275] **you** got it //

Non-standard contraction of preposition and non-standard second-person pronoun [bnatbu02_233]

*NEU: [236] **pr' ocê** ver //

*NEU: [236] *who would've thought of that (lit. **for you** to see) //*

28. Third-person pronoun variant forms

The third-person pronouns received additional reduced forms, as shown in the table below.

Standard and reduced forms of the third-person subject pronoun

Standard form	Reduced form	Translation
---------------	--------------	-------------

ele	e'	he
ela	ea	she
eles	es	plural masculine
elas	eas	plural feminine

The reduced forms may also contract with preposition as provided for by [\[§A.26.3\]](#). The standard norm covers the contraction of the standard form with two prepositions: *de* (of/from), and *em* (in/on/at). The possible contractions of these prepositions with the reduced forms are, thus:

Possible standard contractions with reduced third-person pronouns

Reduced forms	With <i>de</i>	With <i>em</i>
e'	de'	ne'
ea	dea	nea
es	des	nes
eas	deas	neas

Otherwise, the contraction follows the open-list criterion provided for by the special criteria.

Non-standard contraction of preposition with a reduced third-person pronoun [bfamd125_207]

LIA: [207] **pr' ea** tratar //

LIA: [207] **for her** to be treated //

29. Reduced demonstratives

Distal demonstrative forms may be transcribed with a series of reduced forms, as shown in the table below.

Full and reduced distal demonstrative forms

Full form	Reduced form	Translation
aquele	aque'	Sing. masculine distal demonstrative (that)
aquela	aquea	Sing. feminine distal demonstrative (that)
aqueles	agues	Plural masculine distal demonstrative (that)
aquelas	aqueas	Plural feminine distal p demonstrative (that)

The contraction of prepositions and reduced demonstratives follows the same rule in [\[§A.26.3\]](#). The standard norm covering the contractions *a* (to), *de* (of), *em* (in/on/at), the possible forms are:

Possible standard contractions with reduced demonstratives

Reduced form	With <i>a</i>	With <i>de</i>	With <i>em</i>
--------------	---------------	----------------	----------------

aquele	àque'	daque'	naque'
aquela	àquea	daquea	naquea
aqueles	àques	daques	naques
aquelas	àqueas	daqueas	naqueas

Otherwise, the contraction follows the open-list criterion.

Non-standard contraction of preposition and distal demonstrative pronoun [bfamdl20_042]

*OSM: [42] nũ conta nada **p' aque'** (para aquele) cara não /

*OSM: [42] *don't tell it **to that** guy at all /*

Non-standard contraction of preposition and distal demonstrative pronoun [bfammn14_104]

*ANT: [104] **c' aqueas** batidim pesada de' //

*ANT: [104] **with that** heavy walking of him //

30. Diminutive variant forms

Two reduced forms, *-im* (sing.) and *-ins* (plural), are added to the standard diminutive paradigm (*-inho/-inha/-inhos/-inhas*).

Diminutive form [btelpv44_007]

*SIL: [7] tomou **banhozim** agora aí //

*SIL: [7] *you just took a **shower** now //*

31. Pseudo-cleft constructions

31.1 Pseudo-cleft interrogative constructions

In BP, speakers seem to be losing awareness of the presence of the copula verb in cleft interrogative constructions like *que é que* (what is that), *por que é que* (why is that), and *onde é que* (where is that). Pseudo-cleft constructions, where the copula is clearly missing, such as *que que* (what that), *por que que* (why that), *onde que* (where that) are respected in the transcription.

Pseudo-cleft interrogative construction [bnatla04_039]

*ESC: [39] e **como que** cê ficou sabendo disso //

*ESC: [39] *and **how (is) that** you came to know this //*

32.2 Other pseudo-cleft constructions

Other pseudo-cleft constructions may also lack the copula and are transcribed as pronounced.

Other cleft constructions [bpubcv09_377]

*MAR: [377] ela **que** apanha //

*MAR: [377] *she (is the one) **who** gets beaten //*

The standard cleft construction for the example above is *é ela **que** apanha* or *ela **é que** apanha*.

32. Aphaeretic forms

Since aphaeresis may indicate a lexicalization process, aphaeretic forms are transcribed as pronounced. Their occurrences are enlisted on the metadata and inputted to the morphosyntactic parser. Some examples are listed below:

brigado < *obrigado* (thanks)

cabou < *acabou* (it's finished / it's over)

fessor < *professor* (professor)

xá < *deixa* (let/leave)

tendi < *entendi* (I got it)

33. Negation

The reduced form of the negation particle *não* (not/no) is transcribed as *nũ* – not to be confounded with *num*, contraction of *em* (in) + *um* (a). A frequent pattern found in the corpus is the double-negation like in the example below.

Negation [bnatpr09_079]

*ANT: [79] mas **nũ** tá **nãõ** //

*ANT: [79] but it is **not** //

The contraction of the negation particle with *é* (present tense indicative first-person singular variant forms of the verb to be) is transcribed as *n' é* (it is not). The example below exhibits a double negation with a contracted form, and a double negation combined with the negative pronoun *nada* (nothing).

Double negation and contracted form [bnatpr09_079]

*MAR: [40] *é / n' é* barato **não** / viu // [41] **nũ** achei / **nada** barato **não** //

*MAR: [40] *yeah / it isn't cheap / huh* // [41] *I didn't find it / cheap at all* //

34. Variant forms of *senhor/senhora* (Mister/Sir – Mrs./Madam)

The honorifics *senhor* (Mister/Sir) and *senhora* (Mrs./Madam) take on some variant forms, which are respected in the transcription. The following forms are possible:

Variant forms of *senhor/senhora*

Alternative forms	Correspondent standard form
sior	senhor (Mr./Sir)
seu	
sô	
siora	senhora (Mrs./Madam)
sio'	

sá

Below, we give two examples in context.

Variant form of *senhora* (formal second-person pronoun) – bnatla04_059

*HIL: [59] quando eu cheguei do serviço / minha menor virou pra mim e falou / mãe / a / Gabriela falou assim p' **siora** dar uma olhada no computador /

*HIL: [59] *when I got home from work / my youngest daughter was like / mom / (the) / Gabriela asked **you** to check the computer /*

Variant form of *senhor* (Mr.) – bnatps11_005

*CAR: [5] **sô** Geraldo / e a família dele toda //

*CAR: [5] **Mr** Geraldo / *and his whole family //*

35. Intensifier maior/mó

The reduced form of the intensifier *maior* (bigger/very/a lot), *mó*, is respected in the transcription.

Intensifier variant form [bmedex03_207]

*MAR: [207] tá aqui dando **mó** força aqui pra gente //

*MAR: [207] *he's here helping us **a lot** //*