Acoustic gesture modeling. Application to a Vietnamese speech recognition system

Thi-Anh-Xuan Tran

Communauté
UNIVERSITÉ Grenoble Alpes

Institut Polytechnique
de Hanoi

Ce mémoire de thèse est confidentiel

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

préparée dans le cadre d'une cotutelle entre la Communauté Université Grenoble Alpes et l' Institut Polytechnique de Hanoi

Spécialité : **EEATS**

Arrêté ministériel : le 6 janvier 2005 - 7 août 2006

Présentée par

Thi-Anh-Xuan TRAN

Thèse dirigée par **Eric CASTELLI**
codirigée par **Thi-Ngoc-Yen PHAM**
co-encadrée par **Nathalie VALLEE**

préparée au sein de l'**Institut de Recherche International MICA (Multimédia, Information, Communication et Applications) – Hanoi, Vietnam**

et du **Laboratoire GIPSA-Lab (Grenoble Images Parole Signal Automatique) – Grenoble, France**

dans **l'École Doctorale Électronique Électrotechnique Automatique & Traitement du Signal**

ACOUSTIC GESTURE MODELING. APPLICATION TO A VIETNAMESE SPEECH RECOGNITION SYSTEM

Thèse soutenue publiquement le 30 mars 2016,
devant le jury composé de :

Mme. Martine ADDA-DECKER
Directrice de Recherche, CNRS, Laboratoire de Phonétique et Phonologie, Paris, Président

M. Georges LINARÈS
Professeur de l'Université d'Avignon et des Pays de Vaucluse, Avignon, Rapporteur

M. François PELLEGRINO
Directeur de Recherche, CNRS, Dynamique du Langage, Lyon, Rapporteur

M. Eric CASTELLI
Professeur & Chargé de Recherche, CNRS, MICA, Hanoi, Directeur de thèse

Mme. PHAM Thi Ngoc Yen
Professeur, Institut Polytechnique de Hanoi, Co-directeur de thèse

Mme. Nathalie VALLÉE
Chargée de Recherche, CNRS, GIPSA-lab, Grenoble, Co-encadrante

Acknowledgments

Foremost, I would like to express my most sincere and deepest gratitude to my thesis supervisors *Prof. Eric CASTELLI*, *Prof. PHAM Thi Ngoc Yen (MICA_CNRS, Vietnam) and Dr. Nathalie VALLÉE (GIPSA-lab, Grenoble)* for their continuous support and guidance during my PhD program, and for providing me with such a serious and inspiring research environment. A big thank to *Prof. Eric CASTELLI* that guided me throughout all the years of thesis for shaping my thesis at the beginning, for his support, his advice on my research and writing. I am also very thankful to *Dr. Nathalie VALLÉE* for her advice and encouragement during all my thesis period.

I am fortunate to have the opportunity to work with *Prof. René CARRÉ* (DR émérite, CNRS). He taught me various essential knowledge such as speech production, speech perception. I am very grateful to *Prof. René CARRÉ* for his intense participation in the partial orientation of my research.

I highly appreciate the opportunity to know and work with *M. Alexis MICHAUD* (MICA_CNRS, Vietnam). I am sincerely indebted to Alexis for his comments on linguistics and writing.

I would like to very thank to *M. Jean-Marc THIRIET,* director of GIPSA-lab, for accepting me in speech and cognition department. A big thanks to *Prof. PHAM Thi Ngoc Yen* (former director of MICA institute) *and M. NGUYEN Viet Son*, director of MICA institute, for allowing me to work at SpeechCom department.

I take this opportunity to extend my heartfelt gratitude to all members in MICA (especially to the members of the SpeechCom department) and all members in GIPSA-lab (especially to the members of the speech and cognition department), who welcome me to work there and give me a lot of useful comments and discussions concerning my work.

Last but very the importance, I would like to dedicate this moment to my parents and my husband for their endless love and support during all my thesis, who have given me much courage to accomplish this thesis.

ABSTRACT

Speech plays a vital role in human communication. Selection of relevant acoustic speech features is key in the design of any system using speech processing. For some 40 years, speech was typically considered as a sequence of quasi-stable portions of signal (vowels) separated by transitions (consonants). Despite a wealth of studies that clearly document the importance of coarticulation, and reveal that articulatory and acoustic targets are not context-independent, the view that each vowel has an acoustic target that can be specified in a context-independent manner remains widespread. This point of view entails strong limitations. It is well known that formant frequencies are acoustic characteristics that bear a clear relationship with speech production, and that can distinguish among vowels. Therefore, vowels are generally described with static articulatory configurations represented by targets in the acoustic space, typically by formant frequencies in F1-F2 and F2-F3 planes. Plosive consonants can be described in terms of places of articulation, represented by locus or locus equations in an acoustic plane. But formant frequencies trajectories in fluent speech rarely display a steady state for each vowel. They vary with speaker, consonantal environment (co-articulation) and speaking rate (relating to continuum between hypo- and hyper-articulation). In view of inherent limitations of static approaches, the approach adopted here consists in studying both vowels and consonants from a dynamic point of view.

Firstly we studied the effects of the impulse response in the beginning, at the end and during transitions of the signal both in the speech signal and at the perception level. Variations of the phases of the components were then examined. Results show that the effects of these parameters can be observed in spectrograms. Crucially, the amplitudes of the spectral components distinguished under the approach advocated here are sufficient for perceptual discrimination. From this result, for all speech analysis, we only focus on amplitude domain, deliberately leaving aside phase information. Next we extent the work to vowel-consonant-vowel perception from a dynamic point of view. These perceptual results, together with those obtained earlier by Carré (2009a), show that vowel-to-vowel and vowel-consonant-vowel stimuli can be characterized and separated by the direction and rate of the transitions on formant plane, even when absolute frequency values are outside the vowel triangle (i.e. the vowel acoustic space in absolute values).

Due to limitations of formant measurements, the dynamic approach needs to develop new tools, based on parameters that can replace formant frequency estimation. Spectral Subband Centroid Frequency (SSCF) features was studied. Comparison with vowel formant frequencies show that SSCFs can replace formant frequencies and act as “pseudo-formant” even during consonant production.

On this basis, SSCF is used as a tool to compute dynamic characteristics. We propose a new way to model the dynamic speech features: we called it SSCF Angles. Our analysis work on SSCF Angles were performed on transitions of vowel-to-vowel (V1V2) sequences of both Vietnamese and French. SSCF Angles appear as reliable and robust parameters. For each language, the analysis results show that: (i) SSCF Angles can distinguish V1V2 transitions; (ii) V1V2 and V2V1 have symmetrical properties on the acoustic domain based on SSCF Angles; (iii) SSCF Angles for male and female are fairly similar in the same studied transition of context V1V2; and (iv) they are also fairly invariant for speech rate (normal

speech rate and fast one). And finally, these dynamic acoustic speech features are used in Vietnamese automatic speech recognition system with several obtained interesting results.

Key words: vowel gesture, dynamic acoustic features, magnitude of speech, transition direction and rate, SSCF Angles, automatic speech recognition.

Contents

List of figures

List of tables

Abbreviations

ASR	Automatic Speech Recognition
SER	Syllable Error Rate
CV	Consonant-Vowel
VV	Vowel-Vowel
V1V2	Vowel 1-Vowel 2
V1CV2	Vowel 1- Consonant-Vowel 2
C1V1C2	Consonant 1 – Vowel 1 – Consonant 2
V1C2	Vowel 1 – Consonant 2
DP	Dynamic Programming
MFCC	Mel Frequency Cepstral Coefficient
RASTA	RelAtive SpecTrAl
SSC	Spectral Subband Centroid
SSCF	Spectral Subband Centroid Frequency
SSCM	Spectral Subband Centroid Magnitude
F1	The first formant frequency
F2	The second formant frequency
F3	The third formant frequency
Δ	delta parameter
ΔΔ	delta-delta parameter

Introduction

Speech plays a vital role in human communication. Nowadays, with the progress of information and electronics, many advanced machines are present in everyday life, and people want to be able to communicate with these machines by voice. Oral communication is useful for people when their hands are busy; it is also crucial for blind people. We can find many different application fields in real life, such as voice user interfaces, voice dialing (e.g. "Call home"), call routing (e.g. "I would like to make a collect call"), domestic appliance control, search (e.g. finding a podcast where particular words were spoken), simple data entry (e.g. entering a credit card number), preparation of structured documents (e.g. a radiology report), and speech-to-text processing (e.g. word processors or emails), etc.

Speech processing is the study of speech signals and the processing methods of these signals. The signals are now usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signal. Using speech processing methods, human machine interaction can be implemented by human voice. Speech synthesis and speech recognition are two major parts of speech processing. Speech synthesis is the artificial production of human speech. This system converts normal language text into speech. And an automatic speech recognition (ASR) system translates the spoken words into text. The input of an ASR system is the speech signal and the output is the text sequence corresponding to this signal.

Selection of proper acoustic features is perhaps the most important task in the design of a system using speech processing.

Human are able to understand children voices as well as adult voices, male voices as well as female voices, but current ASR systems do not have this ability. That shortcoming comes from the inherent limitations of the recognition method based on the statistical modeling of spectral properties.

For about 40 years, a central (though sometimes only implicit) assumption of automatic speech processing theories is that speech consists of sequences of vowels and consonants are the parts of the speech. In some of the world's languages, while a word can be formed without any consonants, no word may consist of only consonants, without a vowel. The widely used hypothesis is that each vowel constitutes a stable part called "acoustic target". In the human speech production system, those acoustic targets must be reached during vowels production. Therefore, vowels play a very important role in how each word is recognized. But the arrangement of speech in successive phonemes at a phonological level by no means entails that, from a phonetic point of view, speech is made of a

succession of steady states. In fact, there is no such thing as stable state in speech (Greenberg et al., 2002).

Fortunately, in the latter half of the twentieth century, some researchers demonstrated that speech is dynamic process, and natural speech is not a simple sequence of steady-state segments. Some results were given in dynamic speech production and perception and its application (Strange et al., 1983; (W Strange, 1989a); Divenyi et al., 2006; Carré, 2009a; Hermansky, 2011). A spectacular item of research is from Carré's work (2009a) at the DDL-CNRS lab in Lyon, France on the vowel-to-vowel transition. According to his conclusion, surprisingly even when the two target vowels are not actually reached, the produced sound is understood in the same way as when the targets are reached, as long as the direction and the slope of the transition are maintained. He called this "vocalic gestures" or extended to "acoustic gestures".

Following this hypothesis, if it proves feasible to model the dynamic evolution of the speech signal over time as a series of "acoustic gestures", this may be applied into the speech production system as well as ASR system. In the speech production system, the target may not need to be reached. The "acoustic gestures" can be sufficient to indicate the intended target. In the ASR system, we will have the definition of a discriminating criterion, that is, for each identified gesture, the definition of a measure able to recognize it as distinct from all others.

The objective of the dissertation

Therefore we propose this dissertation to try and develop a type of **dynamic acoustic** speech features in **acoustic** system (not in articulatory and/or auditory system) so that they can satisfy some problem as following:

- reliable signal "measurements";
- speaker (gender) independent parameters;
- usable for ASR.

Study context

In my dissertation, we concentrate on analyzing the characteristics of two languages: one is our mother tongue – Vietnamese and one is a foreign language for us – French.

Vietnamese is an Austroasiatic language and it is the national, official language of Vietnam. It is the native language of the Vietnamese (*Kinh*) people, as well as a first or second language for many ethnic minorities of Vietnam (Wiki-English, 2015). Vietnamese is an isolating and tonal language (Trương, 1970; Nathan and Cao, 1988; Mai et al., 1997; Đoàn, 1999). Vietnamese words may consist of one or more syllables (Trương, 1970; Đoàn, 1999). Vietnamese has a complex tone system which is defined not only by the pitch modulation in the syllable, but also by the characteristics of voice quality (ex. vibration of the vocal folds) (Michaud, 2005, 2010; Brunelle, 2009). The number of Vietnamese

tones can vary from six (in the North) to five (speaking in the South), or four at some regional dialects in the Centre (Nguyen, 2004). Each Vietnamese syllable has a unique tone.

According to Wiki-French (2015), French is a Romance language, belonging to the Indo-European family. French is the second-most widespread language worldwide after English, being an official language in 29 countries. It is spoken as a first language in France, southern Belgium, western Switzerland, Monaco, the Canadian province of Quebec and by various communities elsewhere. Differ from Vietnamese, French is the non-tonal language.

These two languages have differences in the possible syllable structures (Trần, 2011). There are also some similarities and some differences in the system of consonants and vowels between two languages. In the scope of our thesis work, we will approach the comparisons between two languages at some similar vowels (such as, /a, ɛ, e, i, u/) based on the proposed acoustic dynamic speech feature which will be presented in this dissertation.

Main task

In order to get results for the above objective, we have to do some experiments to find out firstly the extraction of pertinent acoustic parameters, which could be frequency resonances or frequency spectrum peaks or their pseudo-parameters, probably take into account human perception characteristics (please see in detail in Chapter 3).

And then we try to extract the characteristics of the "shapes" and their representations for phonemes (vowel, vocalic consonants) and also for transitions (vowel-to-vowel, vowel-consonant or consonant-vowel) (please see in detail in Chapters 3 and 4).

After that, we apply all these finding speech features into Vietnamese speech, which presents into some specific characteristics as short vowels or initial/final occlusive consonants produced without burst (please see in detail in Chapters 3 and 4).

Finally, we would like to integrate these features into an automatic Vietnamese speech recognition system to evaluate the effect of these feature models (please see in detail in Chapter 5).

The outline of the dissertation

This thesis includes two main parts: the state of the art (Chapter 1) and our contributions (Chapters 2, 3, 4, and 5).

In Chapter 1, we present firstly the background of speech production, and then review the state-of-the-art on static speech and its limitations and next give the state-of-the-art on speech dynamics, including production dynamics and perception dynamics of speech and its application.

Chapters 2 and 3 bring out our perception experiments. In general, speech signal is completely characterized by its power spectrum and phase spectrum. Phase spectrum is also first candidate of

dynamic speech, but it is also difficult to determine it in natural speech (Yegnanarayana and Murthy, 1992; Alsteris and Paliwal, 2007). On the contrary, the magnitude spectrum is relatively easy to analyze and parameterize into speech features. Therefore, Chapter 2 is devoted to testing the effect of phase spectrum at the speech perception level. Simultaneously, this chapter also addresses the issue whether the amplitude spectrum is clearly sufficient for perceptive discrimination or not.

From the result of Chapter 2, we move on, in Chapter 3, to study dynamic speech, focusing on the influence of the direction and rate of the formant transitions on speech perception. We perform the synthesized Vowel1-Consonant-Vowel2 (V1CV2) speech perception with the absolute frequency values both inside and outside the vowel triangle. We call them no-illusion and illusion experiment, respectively. These two experiments are carried out in order to test the role of the direction and rate of formant transitions on V1CV2 speech perception even when there is no reference to any vowel targets in the vowel triangle.

In Chapters 2 and 3, we conduct the study approach on acoustic dynamic speech feature that still based on the basic components of speech (amplitude/phase spectrum, formant frequencies) with some interesting results; all these parameters are difficult to estimate in natural speech. That is the reason for us to perform our proposed work in Chapter 4 that focus on finding out the new parameters so that they are the similar to formant frequencies, but they can be estimated easily in real speech. And from that point, we propose the way to model the acoustic gestures from the dynamic evolution of the speech signal over time. We propose them as new acoustic dynamic characteristics of speech. However, a question here is raised: whether these parameters are good candidates to distinguish speech. In order to get the answer, we have to do the analysis of these parameters on both Vietnamese and French that are recorded from both their native male and female voice, with both normal and fast speech rate. In each language, we study on the different vowel-to-vowel (VV) transitions and we do comparisons among the obtained results so that we can consider whether these parameters can discriminate different VV speech for each speaker; and whether on the same VV speech, they are variant or invariant among different speakers with both two speech rate. Besides, we also do a comparison between Vietnamese analysis results and French ones with some same VV stimuli in order to consider the similarity and the difference between two different languages at some similar vowels.

While Chapters 2 to 4 evaluate the effect of the acoustic dynamic speech feature by the obtained analysis results and perceptive ones, in Chapter 5 we give another evaluation by their application in Vietnamese speech recognition system. By this way, firstly, we evaluate the performance of the simple dynamic parameters (delta and double delta) that are derived from the static traditional speech features (such as MFCCs). And then, we apply the new acoustic dynamic speech features that are proposed in Chapter 4 into the same Vietnamese ASR.

Our work in this thesis will be ended in Chapter 6 with some discussions from the obtained results at the previous chapters, then we gave some conclusions and perspectives.

Part I. State of the art

Chapter 1
State-of-the-art on speech feature

As a simple model to serve as background to spoken language analysis and speech processing, it can be said that spoken language is used to communicate information from a speaker to a listener. Speech processing does not delve into the cognitive processes involved in thinking and putting words on thought: speech processing mostly cares about the chain of events that begins when stimuli from the brain activate muscular movements to produce speech sounds.

How is speech produced? And, given the considerable gap between the amount of information contained in the signal that reaches the ear, and on the other hand the linguistic units (phonemic, tonal, and intonational) that are retrieved by the listener, which characteristics of speech production suffice for speech perception (perceptual discrimination)?

Section 1.1 will be devoted to reviewing the essentials of speech production to help us understand how human produce speech, also recalling some definitions of speech characteristics, such as: vocal tract, fundamental frequency, formant frequency, etc.

Section 1.2 will review some description on "classical" point of view of speech analysis; and section 1.3 will highlight some problems that stem from this approach.

Section 1.4 will review some dynamic characteristics of speech including production dynamics and perceptual dynamics, as well as applications in speech analysis and speech processing.

1.1 Speech production

All speech begins as a silent breath of air, created by muscular activity in the chest. The air then comes up from the lungs, via the vocal tract, generating a sound wave. This sound is made into speech by various modifications of the supra-laryngeal vocal tract (Robert, 2009), as show on the following figures (Figures 1-1 and 1-2), where the speech organs are divided into three main groups: lungs, larynx which contains the vocal folds, and vocal tract:

Lungs serve as a "power supply" and provides airflow to the larynx.

Larynx modulates the airflow from the lungs and provides either a pseudo periodic sequence of puffs or a noisy airflow source or impulse to the vocal tract. The voice production process occurs at the larynx. The larynx has two horizontal folds of tissue in the passage of air, and they are called the

vocal folds (formerly referred to as 'vocal chords'). The gap between these folds is called the glottis. When the glottis is closed, no air can pass. When there is a narrow opening, that means the vocal folds are held close to each other and oscillate against each other during a speech sound, the sound is said to be voiced. And when the glottis is wide open, as in normal breathing, thus, the vibration of the vocal folds is stopped, and the resulting sound is unvoiced.

Vocal tract is the channel of air flow between the larynx and the mouth and nose. It transforms the sound into intelligible speech. The vocal tract is comprised of the oral cavity from the larynx to the lips and the nasal passage that is coupled to the oral tract by way of velum. The oral tract takes on many different configurations by moving the tongue, teeth, lips, and jaw. When sound is produced at the larynx, that sound can be modified by altering the shape of the vocal tract above the larynx (supra-laryngeal or supra-glottal). The shape can be changed by opening or closing the velum (which opens or closes the nasal cavity connection into the oropharynx), by moving the tongue or by moving the lips or the jaw.

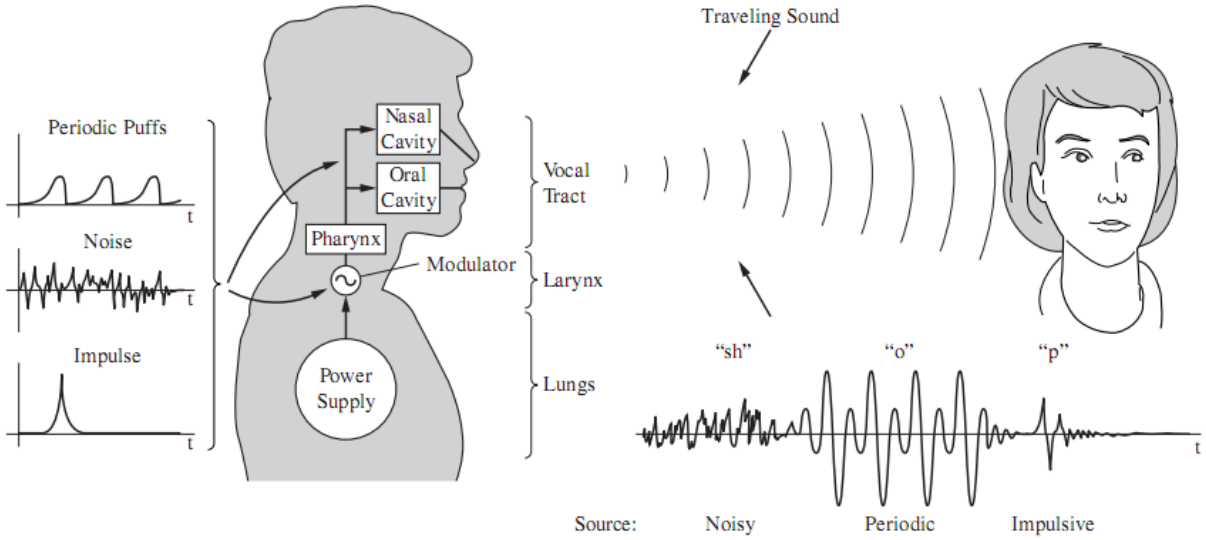


Figure 1-1: Simple view of speech production (Huang et al., 2001).

A consonant is a speech sound which obstructs the flow of air through the vocal tract. Some consonants do this a lot and some do it very little: the ones that make maximum obstruction are the most consonantal, for example, plosives which make a complete stoppage of air stream. Nasal consonants are less obstructive than plosives as they stop the air completely in the oral cavity but allow it to flow through the nasal cavity. Fricatives obstruct the air flow considerably, causing friction, but do not involve total closure. Laterals obstruct the air flow only in the center of the mouth, no the sides, so the obstruction is slight.

Vowels may be classified according to the how far the tongue is from the roof of the mouth during articulation, and how far back in the oral cavity the vowel is articulated.

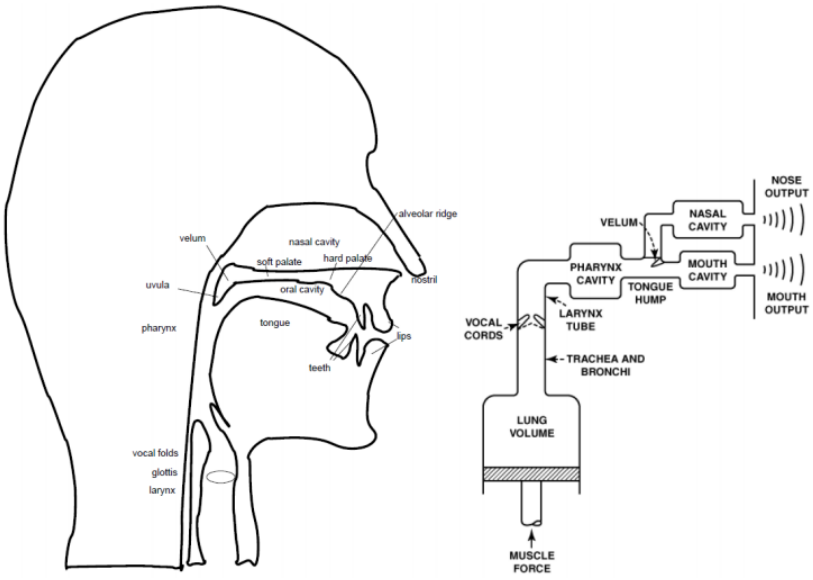


Figure 1-2: A diagram a model of the vocal tract organs: (a) mid-sagittal drawing of vocal organs and (b) a model of vocal tract with discrete components identified (Taylor, 2009).

So, how to model of speech production in acoustic domain?

Speech occurs when a source signal passing through the glottis is modified by the vocal tract acting as a filter.

The process of vowel production can be modeled by a linear combination of the vocal tract as a linear time-invariant (LTI) system, the glottal sound source as an input signal to the system, and radiation characteristics as shown in the following Figure 1-2.

In Figure 1-3, the vowel spectrum P(s) is the product of the spectrum of the glottal source U(s), the transfer function of the vocal tract T(s) (i.e., filter), and the radiation characteristics R(s). This model is also known as "source-filter theory" of vowel production of Fant (1960):

$$P(s) = U(s)\ T(s)\ R(s) \quad \textbf{(1-1)}$$

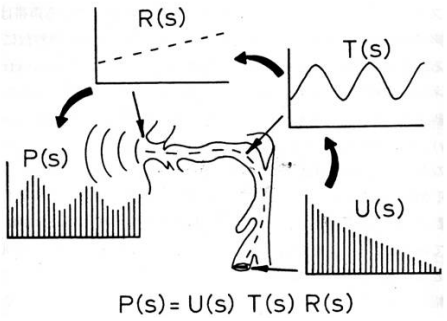


Figure 1-3: Source-filter models of speech production (Fant, 1960) (adapted by Huang et al., 2001).

The rate of cycling (opening and closing) of the vocal folds in the larynx during phonation of voiced sounds is called the fundamental frequency. The audible sound is generated from vocal folds,

includes a lot of simple sound, in which the lowest sound oscillates with fundamental frequency, and other higher sounds oscillate with higher frequencies. These frequencies are the multiples of the fundamental frequencies – these sounds are called harmonics.

There are three general categories of the source for speech sounds: periodic, noisy, and impulsive, although combination of these sources are often present.

Since the glottal wave is periodic, consisting of fundamental frequency (F0) and a number of harmonics (multiples of F0), it can be analyzed as a sum of sine waves. The cavities of vocal tract are excited by the glottal energy. When the shape of vocal tract changes, the resonances also change to form different sounds. The tongue, the jaw, the lips and the soft palate, help shape the oral resonant cavities. Harmonics near the resonances are emphasized, and in speech, the resonances of the cavities that are typical of particular articulator configurations (e.g., the different vowel timbres) are called formant frequencies.

The vocal tract is an open cavity including the nasal cavity, mouth cavity, pharynx and larynx. It extends from the vocal folds to the lips and nostril. The sound resonates in the throat cavity, and mouth cavity. Tongue, jaw, lips and soft palate, help shape the resonant cavity. When the shape of vocal tract changes, the resonances change also, to form different sounds. Harmonics near the resonances are emphasized, and in speech, the resonances of the cavities that are typical of particular articulator configurations (e.g., the different vowel timbres) are called formants.

Speech signals are composed of analog sound patterns that serve as the basis for a discrete symbol representation of spoken language – phonemes (vowels+consonants), syllables and words. And we combine words to make sentences.

Consonants are made by completely or partially blocking airflow during speech. They can be done in different ways: speaker can completely block airflow, push air through a groove or slit to make a hissing sound, block air then make a hiss, in bringing the speech articulators (upper lip, lower lip, jaw, velum, tongue, vocal folds) close together to shape sound. The result is different manners of articulation (different ways of making a sound) (Ladefoged and Disner, 2012).

A question is raised here: which characteristics of speech signal make the sounds distinguishable from one another? Some studies on this topic will be presented in the following section and following chapter.

1.2 State of the art on static speech

In 1952, Gordon E. Peterson and Harold L. Barney published their study of "control method used in a study of the vowels" (Peterson and Barney, 1952). This is a landmark paper in that vowels are generally characterized by the first two or three formant frequencies. Each of them can be represented in the acoustic space (F1-F2 plane) by a dot (Peterson and Barney, 1952). An early acoustic phonetic

study of spectrographic data in terms of manner and place features and of the temporal distribution of information bearing elements appeared in the work of Fant (Fant and Martony, 1962). This specification is static (Chiba and Kajiyama, 1958; Vilain et al., 2015).

Stevens and House (1963) published a study of American English vowels. This paper provided the first outline of a model speech production based on articulatory dynamics. The authors observed that the formant pattern of a given vowel tended to be systematically displayed in the direction of the frequency values of the consonant context. This model was further explored by Lindblom (1963a). The study demonstrated that understood effects in vowels could be predicted from information of vowel duration, consonantal environment and a vowel specific target F-pattern independent of context (Lindblom, 1963a). Lindblom found that there was a direct relation between the duration of a vowel realization and the amount of undershoot as determined from the first three formant frequencies.

A further groundbreaking contribution is the study of Keating (1990), who observed that targets can be described in terms of a 'window model' which replaces positional targets by targets ranges in order to capture under specification phenomena and coarticulation effects (Keating, 1990). Along the same lines, the study of Guenther (1995) showed the DIVA model which describes each phoneme in terms of a set of or-sensory dimensions. These representations are established during a learning stage. They take the form of 'convex region targets' that specify the permissible ranges of variation for each individual dimension. With this revision of the traditional notion of target, Guenther is able to give a unified and compelling account of a number of key phenomena reported in the literature: motor equivalence, the vowel dependence of place of articulation, speaking rate effects including undershoot as well as anticipatory coarticulation.

The mapping of acoustic properties of phonetic segments onto oro-sensory dimensions is supported by empirical work showing that in the absence of auditory feedback speakers are able to maintain auditory goals (Perkell et al., 1993).

All these above speech specifications are static: like Moon and Lindblom (1994), they consider that for each language and each speaker, vowels can be specified in terms of underlying 'targets' corresponding to the context- and duration-independent values of the formants as obtained by fitting 'decaying exponentials' to the data points. The point in focus here is that this specification is static and, significantly, may be taken to imply that the perceptual representation corresponds to the target values (Strange, 1989a,b; Stevens, 2000; Pisoni and Remez, 2005).

1.3 The paradox of static speech approach

From some reviews on static speech on section 1.2, we realize that from the 19th century almost up to the present, the specialists of speech science and speech technology have mostly conceptualized the speech signal as a sequence of static states interleaved with transitional elements intended to reflect the

quasi-continuous nature of vocal production. Henry Sweet is one of the first representatives for this viewpoint (Sweet, 1877). After all, there must be static, stable elements internally if listeners can perceive individual phonemes in the speech stream. While these discrete representation-static targets reached during production and recovered during perception may describe, at best, clearly pronounced "hyper-articulated" speech in which departures from the canonical are rare, it badly fails to characterize spoken language where such departures constitute the norm. A good example for the limitations of phonemic representation is an analysis of 45 minutes of spontaneous conversational speech in which 73 different forms of the word "and" were seen, and yet all of them were unambiguously identified by listeners (Greenberg et al., 2002).

In speech production, steady-states are rare. Vowels can be produced in isolation without articulatory variations, but in natural speech such cases are atypical since their acoustic characteristics are not stable. They vary with the speaker and with the age and gender of the speaker, with the consonantal context (coarticulation), with the speaking rate (reduction phenomena), and with the language (Lindblom, 1963a; Hirtle, 2004; Benzeghiba et al., 2007).

An example can be provided to illustrate this point. The traditional representation of vowels is based on their formants because each vowel has different formant frequencies. Most often, the two first formants, F1 and F2, are enough to distinguish the vowels (Lindblom, 1963a). But the formant values depend on the language, the age and sex of the speaker, etc. Moreover, vowels in fluent speech are rarely steady state on formant frequency trajectories. We can see clearly this on the spectrogram of the sentences *"We have two dogs"* in Figure 1-4 or *"A bell adorns the gate"* in Figure 1-5.

Observe the spectrogram corresponding to the segmentation of phoneme in the top of Figure 1-4 and 1-5 (the boundary of each phoneme is marked by red line), we can see clearly that spectrogram during each phoneme is not stable and it is affected mostly by the contextual adjacent phonemes.

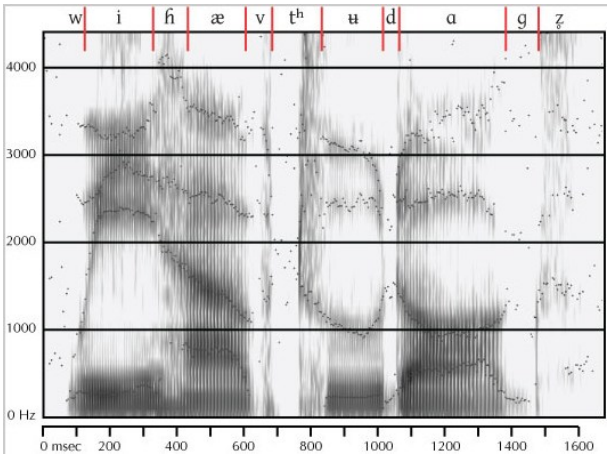


Figure 1-4: Spectrogram of sentence "We have two dogs" (Website 1, 2007).

The example in Figure 1-6 shows obviously for this affectation by observing the formant frequencies of vowel /a/ in English with three words "bab", "dad" and "gag". Obviously, the formant

frequencies of vowel /a/ in three contexts are not the same. So vowels formants do not reach their idealized values because of articulation with adjacent phonemes.

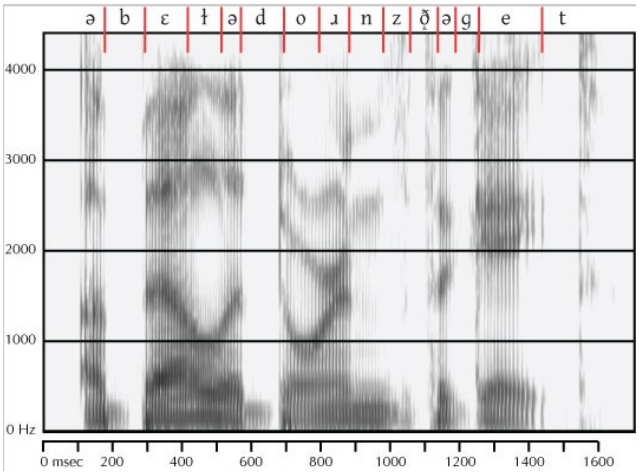


Figure 1-5: Spectrogram of sentence "A bell adorns the gate" (Website 2, 2007).

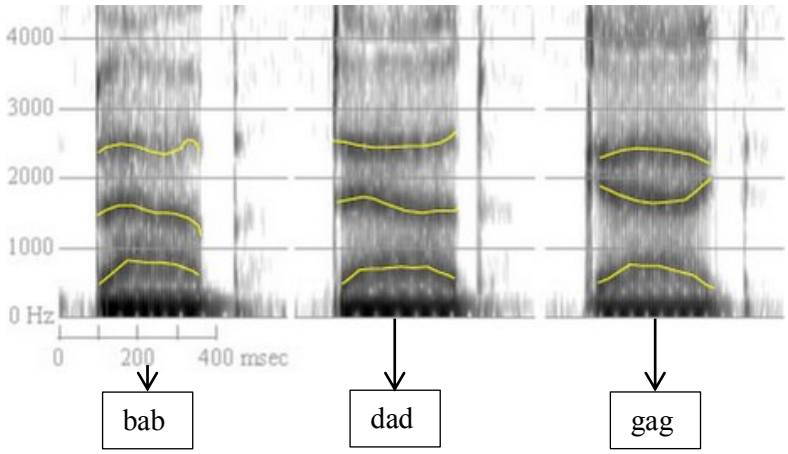


Figure 1-6: Spectrogram of "bab", "dad" and "gag"(Website 3, 2009).

Obviously, we need to supplement phonemic representation by another level that is closer to the phonetics, in order to approach more closely the actual phenomena of verbal communication but in fact it is difficult to determine the separated phonemes in the continuous speech (De Cheveigné, 2003).

At this point, some questions can be raised: (i) how is the perceptual representation obtained if the vowel targets depend on the speaker, and are rarely reached in spontaneous speech production? (ii) are the speech representations the same from one person to another? (Johnson et al., 1993; Carré and Hombert, 2002; Whalen et al., 2004); (iii) how is the vowel/or consonant perceived with its different acoustic characteristics according to the context and the speaker? (Nordström and Lindblom, 1975; Johnson, 1990; Johnson, 1997).

Moreover, sensory physiology indicates that perception is more sensitive to changing stimulus arrays than to static ones (Kandel et al., 2000). Physiology textbooks tell us that sensory mechanisms

have evolved to detect change. In speech perception such processes clearly play a role as shown by investigations of contrast effects in perceptual categorization (Lotto et al., 1997).

So a question here is that if perception likes change, why do systems for phonetic specification favor steady-state attributes?

Some studies to answer those questions will be presented in the next sections and chapters of my dissertation.

1.4 State of the art on dynamic speech

Raymond H. Stetson, a pioneer in the study of speech, wrote that speech is movement made audible (Stetson, 1928). The movements of the speech organs – structures such as the tongue, lips, jaw, velum and vocal folds – result in sound patterns that are perceived by the listener.

A huge challenge for the speech researcher is to find out another way to characterize speech to overcome the limitations of static models, as reviewed in section 1.3 (Lindblom, 1963a; Benzeghiba et al., 2007).

Fortunately, an alternative approach was developed in the latter half of the twentieth century by a team of scientists at the Pavlov Institute of Physiology in St. Petersburg, Leningrad. Headed by Ludmilla Chistovich and Valerity Kozhevnikov, two pioneers of speech science research, this remarkable team recognized that even in clear speech the phoneme could not be considered without the context on which it appeared (Christovich and Kozhevnikov, 1970). Ludmilla Chistovich (1970) said that natural speech is not a simple sequence of steady-state segments. In their view, the phoneme was an epiphenomenon, derived from a more basic unit: the syllable.

Many scientists also advocated and did their research following this way. Firstly, a dynamic specification approach was proposed by (Strange, 1989a,b; Fowler, 1980). In Strange's studies (1989a,b), a series of experiments comes to a successful demonstration that listeners are able to identify vowels with high accuracy although the center portions of CVC stimuli have been removed leaving only the first three and the last four periods of the vowel segment. In other words, vowel perception is possible also on "silent-center" syllables – that is, syllables that lack information on the alleged 'target' but include an initial stop plosive and surrounding formant transitions. Strange took her findings to imply that vowel identification in CVC sequences is based on more than just the information contained within a single spectral slice sampled near the midpoint "target" region of the acoustic vowel segment. Rather the relevant information is distributed across the entire vowel and it includes formant frequency time variations. She wrote "*…vowels are conceived of as characteristic gestures having intrinsic timing parameters* (Fowler, 1980). *These dynamic articulatory events give rise to an acoustic pattern in which the changing spectro-temporal configuration provides sufficient information for the unambiguous identification of the intended vowels*" (Strange, 1989a, page 696).

The term "target" has been used with several different meanings. Strange's definition refers to an observation, the spectral cross-section sampled at a vowel's quasi-steady-state midpoint. It thus differs slightly from the meaning used above, namely, target as a virtual (underlying, asymptote) phenomenon – a point in multi-dimensional articulatory control space that may, or may not, be realized articulatorily and acoustically (Lindblom, 2004).

In the next section, we will review a collection of papers which looks at speech as a dynamic process, including:

- Production dynamics of speech;
- Perceptual dynamics of speech;
- Dynamics in speech applications.

1.4.1 Production dynamics of speech

Firstly, we can refer to the study of Lindblom and his colleagues on the role of gestures in speech and sign language (Lindblom et al., 2006). They examined the concept of "target" in both speech and American Sign Language. In their research, sensory systems prefer time-varying over static stimuli. An example of this fact is provided by the dynamic spectro-temporal changes of speech signals which are known to play a key role in speech perception. To some investigators, such observations provide support for adopting the gesture as the basic entity of speech. An alleged advantage of such a dynamically defined unit – over the more traditional, static and abstract, phoneme or segment – is that it can readily be observed in phonetic records. In their view, the essence of the message contained in each lies in the patterns of movement, rather than in the specific targets to which the motion points. Consonants and vowels cannot be readily distinguished in terms of their articulatory motion; they appear to influence each other.

Secondly, another research concerning this speech dynamics approach must refer to the work of Pols and Van Son (2006). They examined the issue of speech dynamics from the perspective of acoustic primitives (e.g. frequency sweeps) and information theory. They are also concerned with the "efficiency" of communication in term of how the production (as reflected in acoustics) adjusts to specific environmental conditions and communicative tasks. Pols and van Son observed that *"...any vowel reduction seems to be mirrored by a comparable change in the consonant, thus suggesting that vowel onsets and targets change in concert*" (Pols and Van Son, 2006, page 74). This is another way of starting that speech is organized in units larger than the segment. One of the most interesting aspects of their data concerns the importance of context on identification of vowels and consonants. Vowels are generally more accurately identified in full syllabic form. Prosodic factors are also important. In other words, segmental processing is facilitated by syllabic context.

In the context of natural language processing, most ASR systems until now consider that speech can be perceived as a temporal succession of a relatively small number of elementary sounds, called phonemes. But the characteristics of phonemes display considerable variability across speakers and phonemic environments (co-articulation and vowel reduction, etc). (Lindblom, 1963a; Benzeghiba et al., 2007; Carré et al., 2007).

Besides, like every natural system, speech brings its instantaneous state, and its dynamic changes can carry useful information for speech perception. This dynamic approach existed in speech signal processing in its early days, but a lot of researchers ignored it in automatic speech recognition systems. For example, on the topic of speech analysis, we recall the results of Kent and Moll (1969): "*the duration of a transition – and not its velocity – tends to be an invariant characteristic of VC and CV combination*", in which V is any vowel and C is any consonant. The results of Gay (1978) confirmed these observations with different speaking rate and with vowel reduction, namely the reduction in the duration of the vowel, and the transition durations within each rate were relatively stable across different vowels. If the transition duration is invariant across a set of CVs with a constant C and varying Vs, it follows that the transition rate depends on the vowel to be produced. At the very beginning of the transition and throughout the transition there is sufficient information to detect the vowel to be produced. If the perception of the following sound is based on the syllabic duration, on the transition direction and rate, then these results can explain the perceptual outcomes obtained by Strange (1983) in her 'silent center' experiments that replaced the center of the vowel by silence of equivalent duration. She tested manipulation which preserves the direction and the rate of the transition as well as the temporal organization (syllabic rate).

In this way, René Carré and his colleagues developed this theory in French. For instance, Carré and Mrayati (1991) analyzed trajectories in the formant space of vowel-vowel transitions. Their results on natural speech showed that vowel-vowel trajectories in the F1-F2 plane are generally rectilinear. Which suggests that they can be characterized by their direction. More than ten years later, Carré et al. (2007), Carré (2009a) presented one possibility that dynamics can be characterized by the direction and the rate of the vocalic transitions. They said that vowels can be described *dynamically* starting at the very beginning of the transition (Carré et al., 2007).

Carré's hypothesis was studied on French speech production by himself (Carré, 2009a) and was also tested on Vietnamese speech production by Nguyen (2009). The following sections present with some details the results of their studies.

1.4.1.1 Reviewing dynamic characteristic of French vowel-to-vowel trajectories

These studies were referred from René Carré's work on vowel-vowel production (Carré, 2009a). In his experiments, he recorded [V1V2] sequences that were produced by ten French speakers (five males

and five females). Each vowel-vowel item was recorded five times at normal speech rate and five times at fast speech rate.

In [V1V2] sequences, V1 is always /a/ and V2 is one of the French vowels situated on the [ai] ([i, ɛ, e]), [ay] ([y, ø, œ]) or [au] ([u, o, ɔ]) trajectories.

For each [V1V2] sequence, formant frequencies were measured using Praat software each 6.25 ms. The derivation was taken to obtain the formant transition rate. Both formant frequency and formant rate were smoothed.

*1.4.1.1.1 [a**V**] characteristics in the F1-F2 plane*

Figure 1-7 shows the different formant trajectories for [ai], [ae], [aɛ], [ay], [aœ], [aø], [au], [ao] and [aɔ] in the F1-F2 plane for one speaker. Each trajectory is a single production pronounced at normal rate.

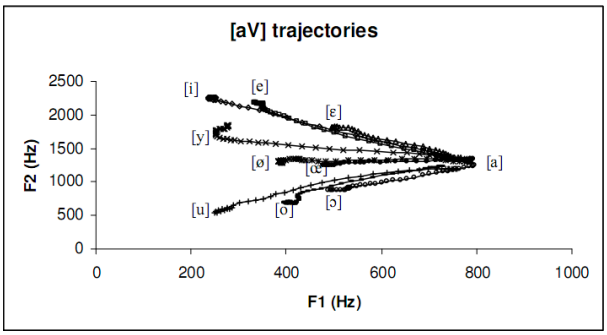


*Figure 1-7: [a**V**] formant trajectories for one speaker (em) in F1F2 plane (normal rate) (Carré, 2009a).*

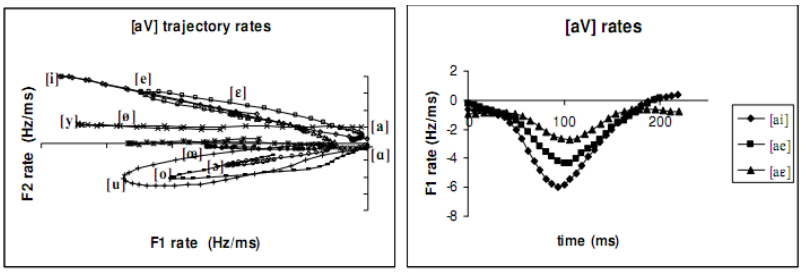
Observing the end parts of the trajectories corresponding to V with the small changes along the trajectory, the results showed that there were three groups of [aV] which can be first discriminated if the directions of the trajectories were taken into account:

- [e], [ɛ] are located along the formant movement of [ai].
- [o], [ɔ] are situated on the [au] trace.
- [ø], [œ] are found on the [ay] trace.

This result was also observed for the other French speakers.

*1.4.1.1.2 [a**V**] transition rate*

Figure 1-8 shows the representation of the F1-F2 transition rate were used to compare the [aV] transitions for the all vowel V:



a) b)

Figure 1-8: [aV] transition rate: a) [aV] trajectory rates in the F1 rate – F2 rate plane for speaker (em); and b) F1 rates in the time domain for [ai], [ae], [aɛ] (Carré, 2009a).

Figure 1-8a shows the results for speaker (em) with one utterance of each [aV]. Figure 1-8b shows the first formant rates in the time domain of [ai], [ae], [aɛ] for one production by speaker (em).

It can be observed that: when [ai], [ae], [aɛ] are compared, the rate trajectory of [ai] is longer than that of [ae] and still longer than that of [aɛ] (see Figure 1-8a). In other words, the maximum rate of [ai] is greater than the maximum rates of [ae] and of [aɛ] (see Figure 1-8b). Thus, the three vowels [i], [e], [ɛ] can be discriminated according to the maximum rates corresponding more or less to the middle of the transition (namely, [ai] maximum rate > [ae] maximum rate > [aɛ] maximum rate). Discrimination can also be obtained throughout the transition and especially from the very beginning of the transition, in other words, from the very beginning of the production task. Similar to this issue, we will see another work in section 1.4.1.2.

Figure 1-9a shows the maximum formant transition rates (mean data and standard deviation for the five productions) in the F1 rate/F2 rate plane for the speaker (em), for both normal and fast production. It clearly shows that there are not large differences between normal and fast productions; and vowels can be discriminated according to their rates.

From the results of sections 1.4.1.1.1 and 1.4.1.1.2, it is clear that the vowels V can be distinguished by the directions and slopes of the [aV] transitions. The transition directions and slopes were considered as the dynamic approach.

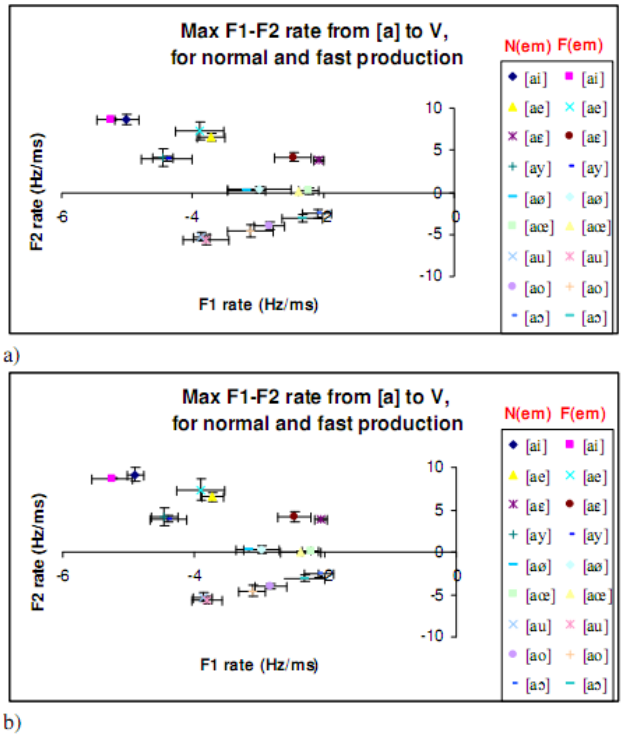


Figure 1-9: a) Vowel transition maximum rates of the transition [aV] for normal (N) and fast production (F) of speaker (em); b) Same data but the formant frequencies F1 and F2 of each [a] vowel at the beginning of the transition are taken into account to normalize the rates (Carré, 2009a).

1.4.1.2 Reviewing dynamic characteristic of Vietnamese speech production

The dynamic approach continued to study on Vietnamese, as in Castelli and Carré (2005), Nguyen (2009). Interesting results were found by Nguyen (2009) on Vietnamese vowels and final consonant /p,t,k/. We will review his results in this section to emphasis the role of dynamic characteristic on discriminating the Vietnamese vowels and Vietnamese final consonant /p, t, k/.

1.4.1.2.1 Vietnamese database (Nguyen, 2009)

In Vietnamese, there exists nine vowels /a/, /ɛ/, /e/, /i/, /ɔ/, /ɤ/, /u/, /ɯ/ and four short vowels /ă/, /ĕ/, /ɔ̆/, /ɤ̆/ (Đoàn, 1999). Only nine long vowels can be pronounced in isolation as on a V syllable. The short vowels cannot appear individually and they must always be combined with a coda consonant.

This Vietnamese vowel corpus was built by four male Vietnamese subjects from the north of Vietnam. Each subject pronounced four following sentences with five times for each sentence:

- "Nói V1C2 êm du" < the meaning in English: "Say V1C2 softly" >.
- "Nói C1V1C2 êm du" < the meaning in English: "Say C1V1C2 softly" >.
- "Nói V1V2 êm du" < the meaning in English: "Say V1V2 softly" >.
- "Nói C1V1V2 êm du" < the meaning in English: "Say C1V1V2 softly" >.

in which, C1 is the initial consonant /b/; V1 is one of the ten Vietnamese vowels (/a/, /ă/, /ɛ/, /ɛ̆/, /ɔ/, /ɔ̆/, /ɤ/, /ɤ̆/, /u/, /i/); V2 is one of the two semivowels /w/, /j/; and C2 is one of the three final consonants /p, t, k/.

The first three formant frequencies of the vowels were measured during all the productions. And then the slopes were compute from their formant transitions.

1.4.1.2.2 The dynamic characteristic on Vietnamese vowel production

A question is raised up here: What characteristics can discriminate Vietnamese vowels (including both long and short vowels), especially each pair of long and short vowels /a/-/ă/, /ɛ/-/ɛ̆/, /ɔ/-/ɔ̆/, /ɤ/-/ɤ̆/? A part in the works of Nguyen (2009) studied on this issues. We will review his results bellow.

Regarding the static approach, for each context, four short vowels [ă], [ɛ̆], [ɔ̆], [ɤ̆] have same target characteristics (on F1, F2, F3) as the four long vowels [a], [ɛ], [ɔ], [ɤ], respectively, but their durations are shorter.

Nguyen compared the dynamic characteristic from the slope of the first three formant frequency for each pair of vowel /a/-/ă/, /ɛ/-/ɛ̆/, /ɔ/-/ɔ̆/, /ɤ/-/ɤ̆/ in the same context, with the final consonants /p, t, k/ and final semivowel /w, j/. The results show that the slope (rate) of the V1C2 or V1V2 transitions to discriminate vowels lie on the same trajectory. This result confirms Castelli and Carré (2005)'s work on production and perception of Vietnamese vowels.

Figures 1-10 and 1-11 below illustrate the case of /a/-/ă/. By observing these two figures, some comments occur: (i) each formant transition slope varies a lots, but the slopes corresponding to long vowel /a/ and short vowel /ă/ for each formant frequency are still different; (ii) the long vowel /a/ and short vowel /ă/ can always be distinguished by at least one formant transition slope of: F1 or F2 or F3. It depends on the final sound, for example:

- The case of the final consonant /p/, the slopes of the third formant F3 are very different.
- Or in the case of the final semi-vowel /j/, the slopes of the two formant F2 and F3 are also very different with no overlap.

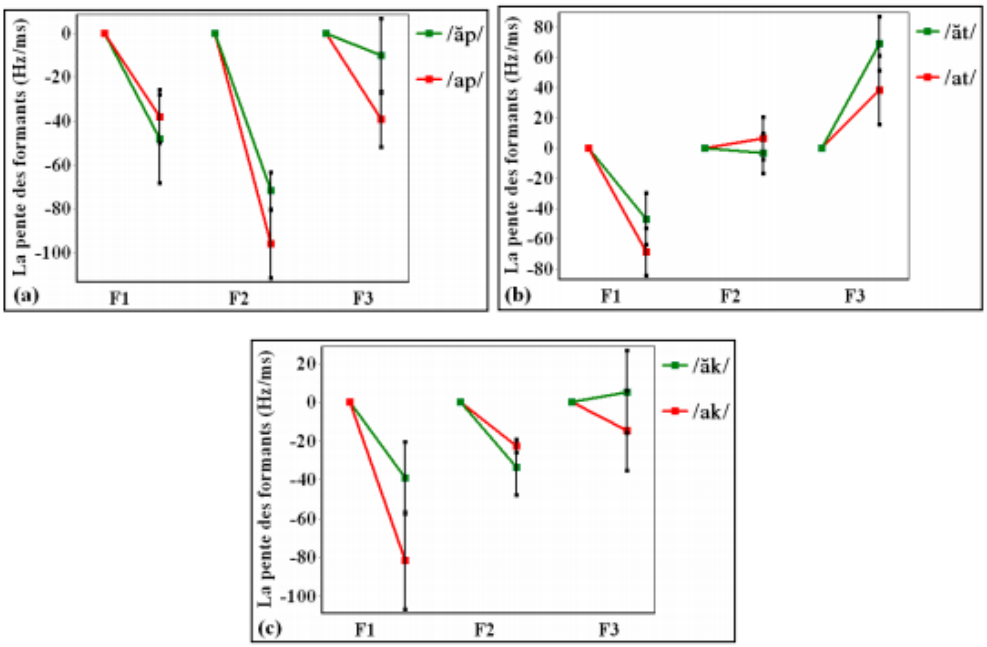


*Figure 1-10: Comparison of the formant transition slopes of F1, F2 and F3 of the two vowels [*a, ă*] in the same context of final consonant: /p/ in (a), /t/ in (b), and //k/ in (c).The slope means and their standard deviation are calculated for all production of V1C2 and C1V1C2 of four speakers (Nguyen, 2009).*

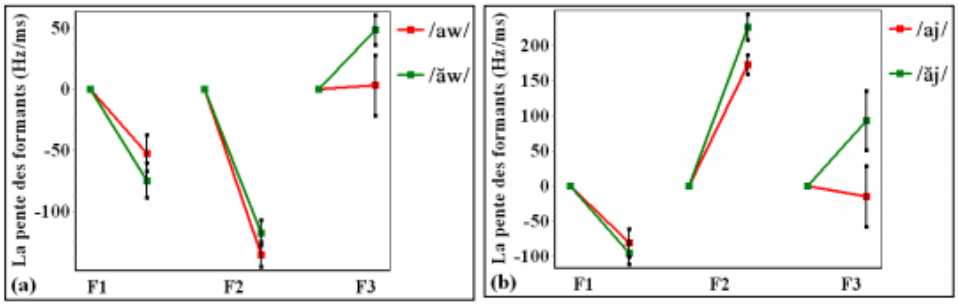


*Figure 1-11: Comparison of the formant transition slopes of F1, F2 and F3 of the two vowels [*a, ă*] in the same context of final semi-vowel: /w/ in a, /j/ in b,. The slope means and their standard deviation are calculated for all production of V1C2 and C1V1C2 of four speakers (Nguyen, 2009).*

These results contribute to affirm again that the slopes of formant transition of V1C2 and V1V2 permit to distinguish acoustically the three pair of the long and short Vietnamese vowels /a/-/ă/, /ɛ/-/ɛ̆, /ɔ/-/ɔ̆/, /ɤ/-/ɤ̆/.

1.4.1.2.3 The dynamic characteristic on Vietnamese final consonant production /p, t, k/

It is well known that both bursts and formant transitions serve as separate perceptual cues to the place of articulation of initial stop consonants. In Vietnamese, final stop consonants (which are voiceless) /p, t, k/ are unreleased (i.e. produced without audible burst). Nguyen et al. (2009) measured the formant slopes of the V1C2 transitions (from the corpus of section *a,*). The results show that in the same preceding vowel context, the three final consonants /p, t, k/ are always clearly differentiated by at least one of the three slopes of F1, F2 or F3.

Figure 1-12 illustrates an example of the comparison of the three final consonants /p, t, k/ in the same preceding vowel contexts /a, i, u/, as follows:

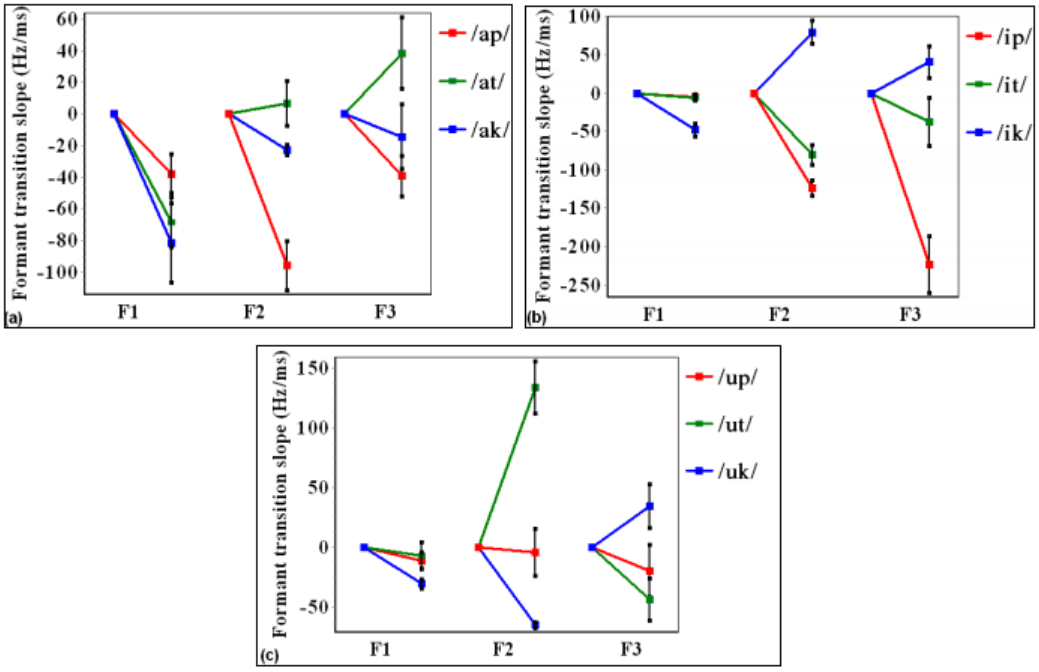


Figure 1-12: Comparison of the formant transition slopes of F1, F2 and F3 of the three final consonants /p, t, k/ in the same context of a preceding vowel: /a/ in (a); /i/ in (b); /u/ in (c) (Nguyen, 2009).

Observed result on Figure 1-12 shows that: (i) depending on the preceding vowel context /a/, /i/ or /u/, the final consonants /p, t, k/ can be distinguished acoustically by at least one of the formant transition slopes F1, F2 or F3; (ii) in the context of three preceding vowels /a, i, u/, the formant transition slope of F2 is always a good cue to differentiate the three final consonants /p, t, k/. This result confirms findings by Serniclaes et al. (2001) about the role of F2 as a cue in the discrimination of the place of articulation for plosive consonants.

1.4.2 Perceptual dynamics of speech

In the previous section 1.4.1, we reviewed the importance and consequences, of considering the dynamic processes behind speech production. But, in order to occur in verbal communication, the perceptual mechanisms must be considered. So in this section, we will review some studies on perceptual dynamics of speech.

In order to assess the dynamic hypothesis of Carré (2009) (presented on section 1.4.1.1), some perceptual vowel studies were conducted focusing on the direction and rate of synthesized transitions. All these work will be review with detail in section 1.4.2.1 below.

Next, another perceptual dynamics of speech will be showed in section 1.4.2.2.

1.4.2.1 Review on Vowel-to-Vowel perception (Carré, 2009a)

In a perceptual experiment, Carré (2009a) focused on the direction and rate of synthesized transitions situated outside the traditional F1/F2 vowel triangle. This situation enables the study of transitions characterized only by their directions and rates without reference to any vowel targets in the vowel triangle of human speech.

1.4.2.1.1 Methodology

In order to test the dynamic hypothesis, René Carré used two first formant frequencies to synthesize four trajectory stimuli outside the vowel triangle (A, B, C, D), as the following Figure 1-13:

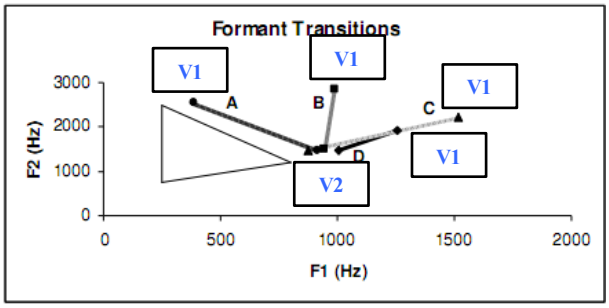


Figure 1-13: The four trajectories (A, B, C, D) in the F1-F2 plane and the vowel triangle. The trajectories are outside the vowel triangle. Their directions and sizes in the acoustic plane vary. (Carré, 2009a).

Figure 1-14 shows the F1, F2 values for the four synthesized stimuli A, B, C and D:

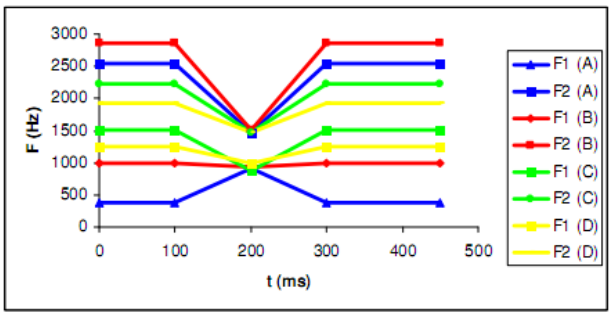


Figure 1-14: F1 and F2 evolutions in the time domain for the four synthetized sequences (A, B, C, D) (Carré, 2009a).

The duration of first part of each sequence was 100ms, the transition was constant and equal to 100ms duration, and the duration of the last part was 150ms. The first and last parts of each sequence were stable and equal in formant frequency; whereas the transitions of the four sequences reached more or less the same point in the acoustic plane (see in Figure 1-14 at around the point of 200ms).

F0 was 300Hz at the beginning of each sequence, held constant during the first quarter of the tonal duration and then decreased to 180Hz at the end.

Twelve French participants (6 males + 6 females) took part in the perceptual test. They were asked to listen to each stimulus and then choose one among five cases proposed, in which there were four

possible responses for each identification test such as [iai], [ɛuɛ], [aua], [aoa]; and the fifth case was noted by "????" symbols that was corresponding to the case of impossible identification (i.e. no response).

1.4.2.1.2 Results in perception and conclusions

a, Results

The sequence A which has the same direction in the acoustic plane and transition rate as [ia] is perceived as [iai]; B which has roughly the same direction and rate as [ɛu] is perceived as /ɛuɛ/; C which has the same direction and rate as [au] is perceived as [aua]; and D which has the same direction of the [au] but is more often perceived as [aoa] at a lower rate.

The results of the experiment are showed in Figure 1-15:

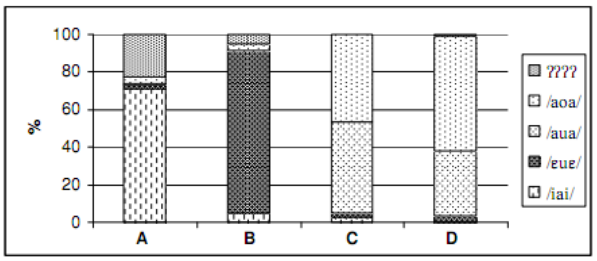


Figure 1-15: Results of the perception test (Carré, 2009a).

The sequence A was identified 71% of the times as [iai]; B was identified 87% as [ɛuɛ]; C and D was identified 95% and 96% as [aua] or [aoa] in which the long trajectory corresponding to a faster rate transition was more [aua] than the short one which was more [aoa]. The option of "no response" was seldom used.

b, Conclusion

The results in perception showed that the region where the four trajectories converge (acoustically close to the low vowel /a/) was perceived either as /a/ or /u/ or /o/ depending on the direction and length (i.e. rate of the transition) of the trajectories. So we can see clearly that at the same position in acoustic plane (that means the same value of F1 and F2), but with different transition directions, different V1V2V1 patterns can be synthesized.

The results obtained by Carré (2009a) suggested also the important role of speech dynamics (namely the direction and rate of the vocalic transitions) in the identification of vowel qualities.

But one issue here is to test Carré's demonstration in vowel-to-consonant-vowel perception. Is this hypothesis still right when an intervocalic consonant is present in the stimuli? This issue will be taken up in Chapter 3.

1.4.2.2 Other previous studies on perceptual dynamics of speech

Carré (2009a)'s study used an original experimental design based on synthetized acoustic transitions to verify the role of dynamics in perception tasks and, in a certain sense, showed perceptual consequences of the dynamic processes of speech production.

Below, we will consider some researches that unveil perceptual aspects of the time-varying acoustic speech signal.

Lublinskaja and colleagues (2006) showed the capacity of amplitude modulation to generate speech from simple non-speech signals. They did perception experiments by asking listeners to identify synthesized speech-like stimuli. The stimuli were vowel-like sounds consisting of "formant impulses" (short tonal impulses with triangular envelop shape, which approximate the output of a band-pass filter tuned to a frequency of F1, F2 or F3). A three-formant impulse was synthesized by adding the waveforms of the F1, F2 and F3 impulses with equal coefficients. The impulse duration was 10ms, which corresponds to a fundamental frequency of 100 Hz. The stimuli contained 25 formant impulses, so that the total duration of a stimulus was 250ms. The frequencies of formants were chosen in order to make the stimuli similar to the vowel /a/, as well as making the distances between the formants approximately equal to each other on the bark scale, such as F1 = 0.6kHz, F2 = 1.2kHz, and F3 = 2.2kHz. The signals were synthesized at a 20kHz sampling frequency and a low-pass filter cut-off of 10 kHz, using a 12 bit digital-to-analog converter.

Two types of stimuli were used in these experiments:

- Type 1: the amplitude of one of the formants (F1, F2 or F3) or of all the three formants (F1, F2 and F3) was attenuated during the initial 80ms of the sound. The amount of attenuation could vary from 0 to 40dB compared to the initial level. The amplitude envelope events are the onset and offset of a segment, referred hereafter as the on- and off-events, for short. There are four versions of quasi-syllables in their experiments: for three versions, the on-event is located in one of the three possible frequency regions. In the fourth version, the on-event appears simultaneously in the three regions. Figure 1-16 shows the spectrograms for all four versions of quasi-syllable and illustrates temporal changes in formant amplitudes while formant frequencies remain unchanged.
- Type 2: Stationary vowel-like stimuli build with constant formant amplitudes during the whole stimulus duration.

For each type, formant frequencies did not vary in any of the cases, providing no frequency transitions in the stimuli.

Nine listeners without hearing impairment participated in the experiments. They were asked to identify each stimulus either as a vowel, or as a syllable consisting of a consonant following a vowel, or as a vowel with some irregularity.

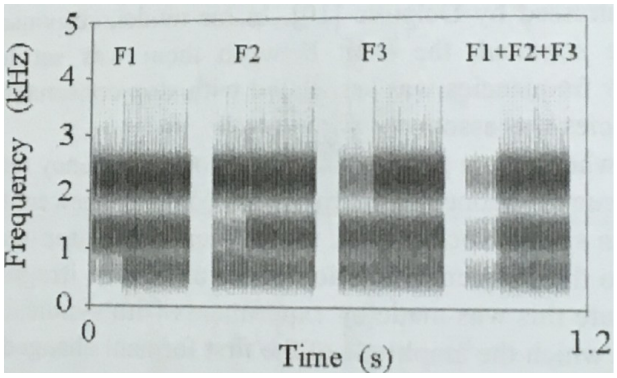


Figure 1-16: Spectrogram of stimuli in the experiment with attenuation, during the initial segment of sound, of one or several maxima in the spectrum. In the stimuli from left to right, either F1, F2, F3 or all three formants are attenuated, respectively (Lublinskaja et al., 2006).

Results of using stimuli of type 1 and 2 are shown below in Figure 1-17, which presents the distribution of responses (identification score) in relation to the formant amplitudes of the stimuli. The data were pooled for all levels of amplitude attenuation during the initial segment of the sound. Responses are grouped into the following categories:

- A vowel (of whatever phonetic quality). The stimuli were most often identified as /a/ but responses /e/, /i/, /o/ and /y/ also occurred. All responses consisting only of a vowel were combined into a single category represented as a V in the leftmost column in Figure 1-17.
- A vowel (of whatever phonetic quality) with some amplitude irregularity, indicated as V~ in Figure 1-17.
- A syllable consisting of an initial lateral consonant /l/ followed by /a/, indicated as LA.
- A syllable consisting of a nasal consonant /m/ and a vowel /a/ indicated as MA.
- A syllable [na], indicated as NA.

The results contained in Figure 1-17 show that the stationary vowel-like stimuli were mostly identified as vowels, while the stimuli with an amplitude jump after the initial 80ms of the sound were perceived as syllables. The quality of a consonant in the perceived syllable seems to be determined by the frequency position of the amplitude jump. Attenuation of F1 results in identification of the syllable as /ma/, and attenuation of F2 leads to identification as /la/. Attenuation of F3 results in a more heterogeneous distribution of responses which are divided between a vowel and the /la/-syllable. The widest dispersion of responses was observed for stimuli with the amplitudes of all three formants attenuated at the beginning of the sound.

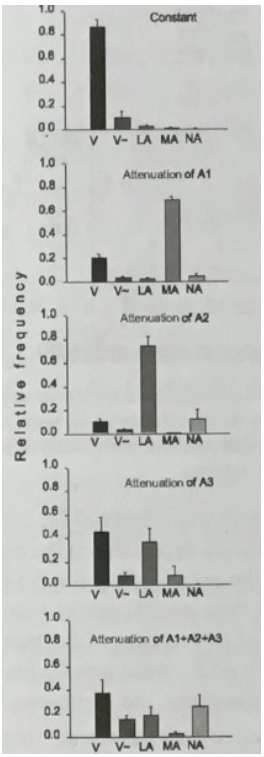


Figure 1-17: Distribution of responses in the identification experiment. The panels represent, from top to bottom, responses to the stimuli with no amplitude attenuation; attenuation of F1, F2, F3; or of all three formants in the spectrum respectively (Lublinskaja et al., 2006).

In the first series of experiments, Lublinskaja et al., (2006) reported that attenuating the amplitude of one, or more, formants had an effect on the identification of the quality of the second segment (i.e., the vowel) in a quasi-syllable, despite its spectral position remaining constant for all categories of stimuli. The experiments demonstrated that amplitude jumps at different frequencies yields specific phoneme qualities of consonants, and an abrupt change in the amplitude of one of the formants may influence the perceived quality of a following vowel.

Related to the topic of dynamic perception, the presence of a rapid frequency transition, in addition to conferring phonetic identity to speech sounds, also helps localizing its source in the horizontal plane, as shown by the research of Richard M. Stern and his colleagues in (Stern et al., 2006). Without amplitude and frequency modulations in the speech signal, a listener would find it more difficult to segregate a target talker's speech from the babble produced by voices of a crowd.

Richard Turner and his colleagues (Turner et al., 2006) presented another interesting dimension of perceptual speech dynamics. They showed that temporal processing in the ear, – in addition to the short-term analysis used to extract periodicity information for pitch and the long-term analysis to extract syllabic and sub-syllabic segments –, also keeps track of the duration and shape of the resonance patterns inside each separate auditory channel. That information is indispensable for characterizing the speaker's gender, age and size – in other words, the speaker's identity – and offers

explanation of why different formant values across men, women, and children are nonetheless understood as belonging to the same vowel category.

1.4.3 Dynamics in speech applications

Obviously, as shown by results of previous studies presented in sections 1.4.1, and 1.4.2, speech is a dynamic process. In this school of thought, some researchers in the world recently started to work on how to incorporate speech dynamic aspects into real applications.

We propose below to survey different approaches of modeling using speech dynamics. Among them, the first is certainly Furui (1986)' s study on using dynamic features of speech spectrum into a speaker independent isolated word recognition system. Spoken utterances were represented by time sequences of cepstrum coefficients and energy. Regression coefficients for these time functions were extracted for every frame over an approximately 50ms period. Time functions of regression coefficients extracted for cepstrum and energy were combined with time functions of the original cepstrum coefficients, and used with a staggered array dynamic programming (DP) matching algorithm to compare multiple templates and input speech. Regression analysis was applied to each time function of the cepstrum coefficients and to the log-energy over several frames every 8ms involving the following calculation of the linear regression coefficient, namely, the first-order orthogonal polynomial coefficient:

$$a_m(t) = \left(\sum_{n=-n_0}^{n_0} x_m(n) \cdot n\right) \Big/ \left(\sum_{n=-n_0}^{n_0} n^2\right) \quad \textbf{(1-2)}$$

where $x_m(-n_0 \le n \le n_0)$ is the time function of the m^{th} parameter within the segment being measured; $x_0(n)$ is the log-energy and $x_m(1 \le m \le 10)$ is the m^{th} cepstrum coefficient, which represent the slope of the time function of each parameter in each segment respectively, and called dynamic features of speech spectrum. The utterance is then represented by time function of the log-energy $x_0(t)$, cepstrum coefficients $\{x_m(t)\}_{m=1}^{10}$, and the regression coefficients $\{a_m(t)\}_{m=0}^{10}$, where t is the frame number. Speaker-independent isolated word recognition error rate of 2.4 percent could be obtained with this method. Using only the original cepstrum coefficients, the error rate was 6.2 percent. Obviously, the technique based on combination of instantaneous and dynamic features of speech spectrum was shown to be highly effective in speaker-independent speech recognition.

Another approach is the study of Atlas (2006). He directly addressed dynamic aspects of speech by characterizing the modulation spectrum of speech. Atlas was not interested in spectral slices, instead he wanted to measure how the information in auditory bands changes over time. Altas measured the spectrum of the slow modulations in each channel (a modulation spectrogram). This information was useful to separate two speakers based on their separate modulations.

Hynek Hermansky also introduced the important role of spectral dynamic using Relative spectral (rasta) filtering to improve the performance of ASR (Hermansky, 1994; 2011). Hermansky noticed that performance of even the best current stochastic recognizers severely degrades in an unexpected communication environment. In some cases, the environmental effect can be modeled by a set of simple transformations and, in particular, by convolution with an environmental impulse response and the addition of some environmental noise. Most of the time, the temporal properties of these environmental effects are quite different from the temporal properties of speech. Hermansky experimented with relative spectral filtering approaches that attempted to exploit these differences to produce robust representations for speech recognition and enhancement. He described the convergence between successful statistical approaches to speech recognition and that which is already known about human perception: the relative insensitivity of human hearing to slowly-varying stimuli may partially explain why human listeners do not seem to pay much attention to a slow change in the frequency characteristics of the communication environment, or why steady background noise does not severely impair human speech communication. However, even when the experimental evidence from human perception may give only limited support, the suppression of slowly-varying components in the speech signal makes good engineering sense. Thus, to make speech analysis less sensitive to the slowly changing or steady-state factors in speech, Hynek Hermansky used a spectral estimate in which each frequency channel is a band-pass filtered by a filter with a sharp spectral zero at the zero frequency. Since any constant or slowly varying component in each frequency channel is suppressed by this operation, the new spectral estimate is less sensitive to slow variation in the short-term spectrum. This filter is an IIR filter with the transfer function:

$$H(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \quad \textbf{(1-3)}$$

The low cut-off frequency of the filter determines the fastest spectral change of the log spectrum, which is ignored in the output, whereas the high cut-off frequency determines the fastest spectral change that is preserved in the output parameters.

This new features were used as input to a conventional hidden-Markov model recognizer and the results (Hermansky, 1994) showed that the rasta features can improve the recognition accuracy in noise environment.

Vicsi's work (2006) considered the dynamics of speech for a different purpose for training students to speak better. She presented a computer-assisted language learning system explicitly based on dynamic changes in the speech waveform. Her work helped students improve their speech by visualizing different aspects of the speech they produce.

1.5 A first study on acoustic Vietnamese vowel gesture based on formant

The first study on acoustic vowel gesture was performed by Alliot (2009). This gesture system was based on formant F1 and F2 frequencies of Vietnamese vowel transitions.

1.5.1 Methodology

The estimation of formant frequencies is necessary for performing the acoustic vowel gesture on formant F1-F2 plane. In order to estimate formant frequencies for Vietnamese vowels, contrary of some authors that used available software toolkit (Castelli and Carré, 2005; Nguyen, 2009), Alliot (2009) proposed an algorithm to estimate two first formant frequencies automatically. He proposed one algorithm to estimate the two first pseudo-formant frequencies. This algorithm was illustrated in Figure 1-18, as follows:

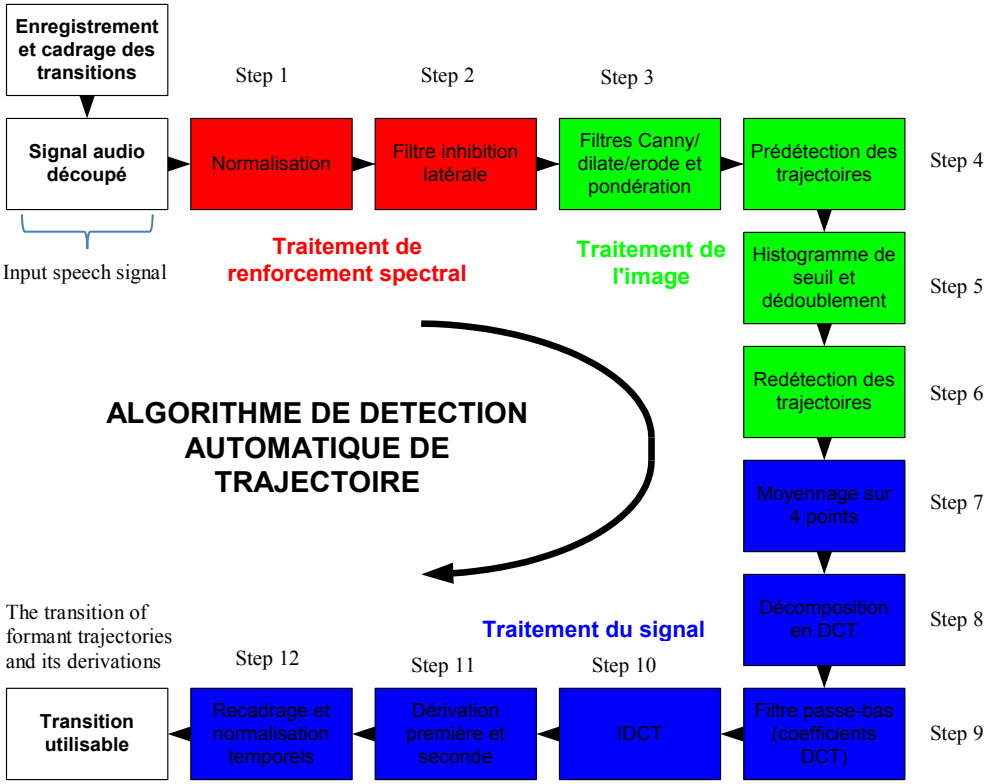


Figure 1-18: Algorithm diagram for estimating the two first formant frequencies of Vietnamese vowels (Alliot, 2009).

Alliot's algorithm presented in Figure 1-18, provides basically an approach to estimate the two first formant frequencies F1 and F2.

Firstly, speech signal will perform normalization on time domain after framing (step 1) and then it is pushed into a lateral inhibition filter (step 2). The lateral inhibition is able to increase temporal and spectral irregularities. This means:

- On frequency domain, the lateral inhibition filter enhances the spectral peaks. This can be used to detect changes in formants;
- On time domain, it shows temporal contrast amplification and hence it can highlight stable regions of vowel speech signal.

Next, the signal will be continued to pass the block of "Filtres Canny/dilate/erode et pondération" (step 3) in order to filter and detect formant frequency contours.

Before the duplication step "Histogramme de seuil et dédoublement" (step 5), the trajectory of two first formant frequencies is cut off (step 4) by lack of continuity in the points, and then a trajectory known to be empirically and statistically good can be detected.

After the process "Histogramme de seuil et dédoublement", the trajectories of the two first formant frequencies are detected one again (step 6).

The obtained trajectories will push into from step 7 to step 10 to smooth the two first formant trajectories. After these steps, the used results of the two first formant trajectories are obtained.

In step 11, the derivations of the two first formant trajectories are computed.

Finally, in the final step (step 12), the formant trajectories and its derivations trajectories are cropped in time domain to keep the transition.

After the twelve steps, formant transition trajectories and their derivations will be plotted in F1/F2 plane and F1 velocity (or speed)/F2 velocity (or speed) plane, respectively.

1.5.2 Stimuli

Alliot (2009) studied six Vietnamese vowel-to-vowel stimuli /ai, ia, au, ua, ui, iu/. Each item was recoded three times by a native Vietnamese male speaker from the north of Vietnam with age of 32 years old. The two first formant frequency trajectories and their derivations were obtained by the algorithm described in section 1.5.1.

1.5.3 Results

The left and right sides of Figure 1-19 shows the two first formant frequencies trajectories on the F1/F2 plane and the formant speed trajectories on the speed plane, respectively, of the six vowel-to-vowel pattern /ai, ia, au, ua, ui, iu/ for one native Vietnamese male speaker.

The observed result in Figure 1-19 shows that firstly in the F1/F2 plane, all formant trajectories are fairly straight lines. According to the direction of these trajectories, it can be visually distinguished visually three groups of trajectories:

- /ai/ and /ia/ trajectories are presented by solid/dashed blue lines, respectively;
- /au/ and /ua/ trajectories are plotted by solid/dashed green lines, respectively.

- /ui/ and /iu/ trajectories are plotted by solid/dashed red lines, respectively.

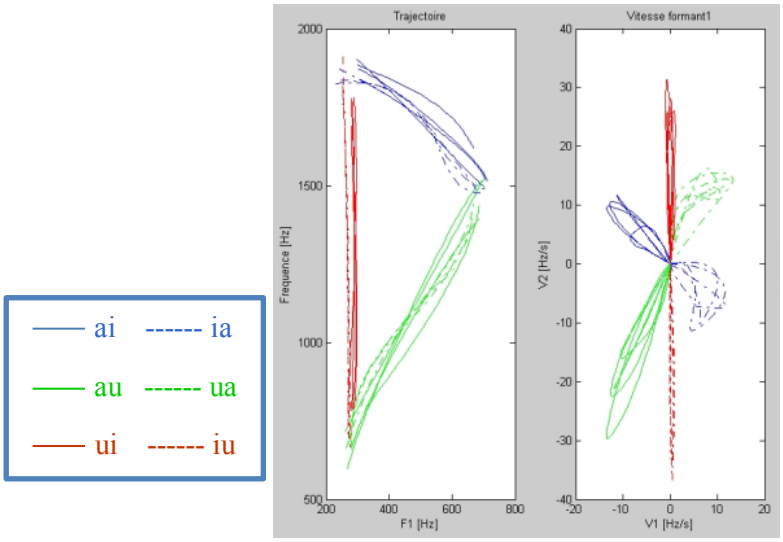


Figure 1-19: Vowel-to-Vowel transition in the plane F1/F2 (left) and in the F1-speed/F2-speed plane produced by a native male speaker of Vietnamese (Alliot, 2009).

Secondly, in the speed plane, the speed trajectory of each transition has more or less ellipse shape, and obviously, can be completely separated from others. This result allowed Alliot to conclude that the six transitions of /ai, ia, au, ua, ui, iu/ can be distinct completely, since the speed direction of each transition is fairly opposite to one of its inverse transition (example for /ai/ and /ia/; or /au/ and /ua/; or /ui/ and /iu/).

From Alliot's results, we ourselves realize an interesting point that the acoustic gesture can be defined with simple form: the straight line of formant frequencies and the ellipse shape of its formant speed. Therefore, if we can characterize the formant trajectories and its speed, we can describe acoustic gesture.

1.5.4 Limitations

Alliot's result was just presented for a Vietnamese male voice. This raises the following issue: what happens with Vietnamese female voice? May we obtain the same acoustic gesture from Alliot's algorithm for female voice? This issue is still big traditional for researchers on formant measurements.

Moreover, the algorithm is very complex, which increases the computation cost.

1.6 Conclusions of chapter 1

Chapter 1 on state of the art of speech feature presented firstly a brief review of static approaches to speech descriptions of a "classical" point of view of speech specification. However, this approach faces many limitations that were presented in more details in section 1.3.

Therefore, in order to solve these problems, some studies (reviewed in sections 1.4.1 and 1.4.2) proposed a new point of view that all speech production is a dynamic phenomenon. The perceptual

mechanisms responsible for decoding the speech signal must also take into consideration its dynamically changing nature.

Carré's studies on speech production (Carré, 2009a; Carré et al., 2007) gave the invariant hypothesis on dynamic acoustic of speech, and then used perception data as a way to verify production dynamics and the results showed perceptual consequences of the dynamic processes of speech production.

In recent years, some researchers in the world started thinking about how incorporate the dynamic aspects of speech into real applications (Hermansky, 2011; Divenyi et al., 2006). Some different approaches to modeling and using the dynamics of speech were presented in section 1.4.3.

From the above results, we can see obviously the dynamic characteristics of speech production and speech perception. Some remarkable points here are:

- The role of speech magnitude is important in speech intelligibility.
- The role of transition direction and length on frequency domain, and of transition rate (slope) in time domain are important to distinguish VV, CV stimuli.
- Some researchers applied the speech dynamics in speech applications and all obtained good results. One of the outstanding results is about spectral dynamics. This feature improved the accuracy of speech recognition system (Hermansky, 2011).

  Although these above dynamic parameters are still based mainly on spectral domain and then convert to cepstral domain, and the dynamic concept here is chiefly the derivative and acceleration from the static parameters, but these parameters also premised firstly for the dynamic speech feature approach in real speech application in general, and in automatic speech recognition system in particular.

We indeed realize the importance role of dynamic speech on speech processing, and I myself inherit and continue to develop some studies on dynamic speech feature, after that I apply them into Vietnamese recognition system.

Part II. Contributions

Chapter 2
A study of speech signal in terms of amplitude and phase

The studies we reviewed in Chapter 1, pointed out the dynamic characteristic in speech production and speech perception and showed the important role of dynamic speech on speech application. From Chapter 2 to Chapter 6, we will propose some new studies that still follow closely the ***dynamic*** speech feature on ***acoustic*** domain.

As we know, speech signal is completely characterized by its power spectrum and phase spectrum (Paliwal and Alsteris, 2003). Phase spectrum is also a first candidate of dynamic speech, but it is also difficult to determine it in natural speech (Yegnanarayana and Murthy, 1992; Alsteris and Paliwal, 2007). On the contrary, the magnitude spectrum is relatively easy to analyze and parameterize into speech features. Therefore in Chapter 2, we propose to test the effect of phase spectrum at the speech perception level. Is it important in perceptual speech? Is it a good candidate as a feature to characterize the dynamics of speech? Concurrently, the subsequent sections are also dedicated to addressing the issue of the role of speech amplitude in acoustic-perceptual discrimination.

2.1 Introduction

Formant frequencies are considered as the traditional acoustic characteristics directly link to speech production. With vowels, the frequencies of formant determine which vowel you hear and, in general, are responsible for the differences in quality among periodic sounds. Although it is usually assumed that female voices are as intelligible as male voices, but in fact, it is well known that speech formant frequencies is very difficult to estimate accurately in automatic speech processing system, especially in the cases involving high fundamental frequency voices as female or child voices. Because of high fundamental frequency, the harmonic spectrum tends to be less dense with more widely spaced harmonics and the formant frequencies cannot be well defined and detected. Figure 2-1 is an example for this problem showing the sonogram of a CVC Vietnamese sequence /bɤ̆k/ pronounced by a

Vietnamese female voice (with F0 around 300 Hz). Looking at the spectrum of this female voice, we see only harmonics, the formant structure disappears because of some harmonics present.

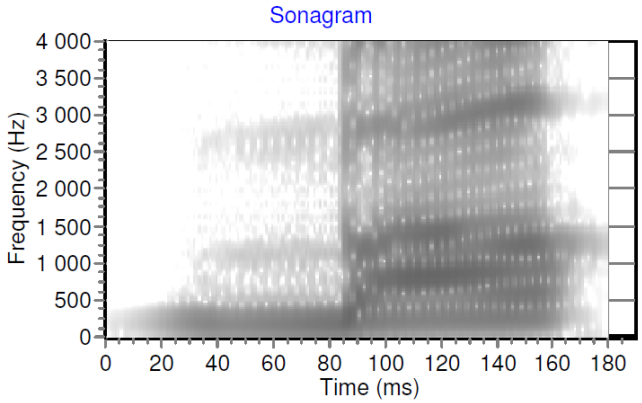


Figure 2-1: Sonogram of /břk/ pronounced by a Vietnamese female voice.

Some authors have shown that dynamical aspects of Vietnamese short vowels are important in perception (Castelli and Carré, 2005; Nguyen, 2009; Nguyen et al., 2009). In order to distinct vowels (both long and short Vietnamese vowels) in CV context, in their experiments, they obtained formant trajectories of natural CV productions by male voices and then calculated formant slopes (rates). And when they took into account the slope (rate) of the CV transitions, the vowels can be discriminated according to the maximum slopes of the positive peaks. The formant slopes here were the dynamic acoustic parameters. But here, the problem in determining where is the formant transition segment (the beginning and the end of the transition) in all CV speech signal is still a big challenge.

However speech is a non-stationary signal where amplitude, fundamental frequency and formant frequencies vary with time. During any of these transient states, the speech signal contains information on the whole vocal tract transfer function. It seems interesting to test the importance of these transients at the perception level.

So what parameter(s) is (are) used by the human auditory system (perception) to perceive male voices (with rich harmonic spectra) and female and child voices (with poor harmonic spectra)? From the above survey, we intend to study the importance of the vowel amplitude transients at the perception level, in which:

- Impulse response at the beginning and at the end of the production: the spectrum of the impulse response is the transfer function whatever F0. F1 and F2 are clearly defined but only at two points in the time domain. Experiment 1 was designed to verify that impulse responses are also observed in the beginning and at the end of the vowel signal with poor harmonic spectra.
- Impulse response during the transition: in the case of a tube, any deformation brings impulse response (and thus information on the whole spectrum) whatever F0. So theoretically, if this

information is processed by the auditory system, the V1V2 transitions could be perceived. This problem was studied in experiment 2.

On another side, the vocal tract is a system with minimum phase: it means that there is a strict correspondence between the amplitude of the transfer function and the phase. This characteristic of the vocal tract system is similar for oral vowel production. It is also well known that different signals with stable different phases between harmonic components are perceived as the same (Huggins, 1952), but what about signals with dynamic changes of phases (leading to frequency variations)? Although it was commonly assumed for decades that most of the speech signal information is transmitted by amplitude variations, several studies have shown the importance of phase variations in speech intelligibility. From an analysis using the speech re-synthesis technique, followed by perception tests, Alsteris and Paliwal (2003) studied the perception of /aCa/ sequences where C is one of the 16 main consonants of Australian English (Paliwal and Alsteris, 2003). They showed that the phase variations appear to play an equally important role as the magnitude variations in the intelligibility of consonants. However, they could not explain the nature of the involved phenomena leaving the question on the role of phase variations in speech perception unanswered.

So, what is the effect of phase spectrum at the speech perception level? Is it important in perceptive speech? Is it a good candidate of the dynamic speech feature? Concurrently, we would like to assess the role of speech amplitude in acoustic-perceptual discrimination. For this purpose, experiments 3 and 4 were conducted:

- Experiment 3 examined the characterization of speech signal from its power spectrum and phase spectrum, and that speech signal is completely characterized by them.
- Experiment 4 was proposed to test the role of amplitude spectrum in speech perception without phase information.

We intend to perform all four experiments by synthesis, because we want to reproduce, with naturalness, even with specific sounds of Vietnamese language, as the short vowels and final consonants without burst.

In the experiment 4, we focus only on short vowels with female voice. The Vietnamese short vowels cannot be pronounced in isolation as long ones, so they cannot appear individually and they must always be combined with a code. And then perceptual comparisons between synthetic /bVk/ will be set up, in which V is one of the Vietnamese short vowels /ɤ̆, ɔ̆, ă/.

All speech signal stimuli in our four experiments were synthesized with high fundamental frequencies (i.e. female voices) which have poor harmonic spectra.

2.2 Characteristics of impulse response and magnitude of the spectral components in Vietnamese speech

2.2.1 Experiment 1 – Impulse responses are produced in natural speech

2.2.1.1 Methodology

Vowels are generated by a formant synthesizer. The formant synthesizer function corresponds to the one of the vocal tract. And in speech production, the amplitude control is placed before the vocal tract.

In this experiment, we would like to verify whether impulse responses are observed in the beginning and the end of speech signal. Therefore, we perform two formant synthesizer of vowel that correspond to two positions of the amplitude circuit, as follows:

- The first case, the amplitude circuit is placed between the formant circuits and the pulse generator F0. The performance diagram of this experiment is given in the Figure 2-2.
- The second case, the amplitude control is placed between the formant circuits and the output speech. This experiment is described in the Figure 2-3.

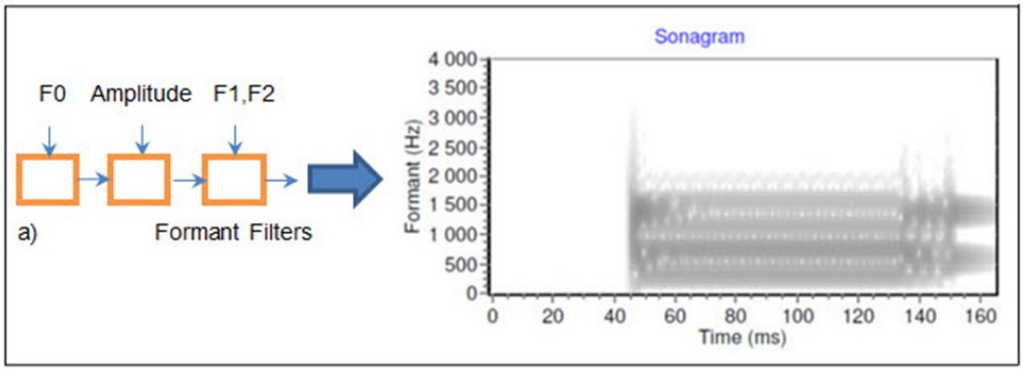


Figure 2-2: Formant synthesizer: the amplitude gate is placed before the formant circuits; and its sonogram of synthetic /a/, F0=400Hz.

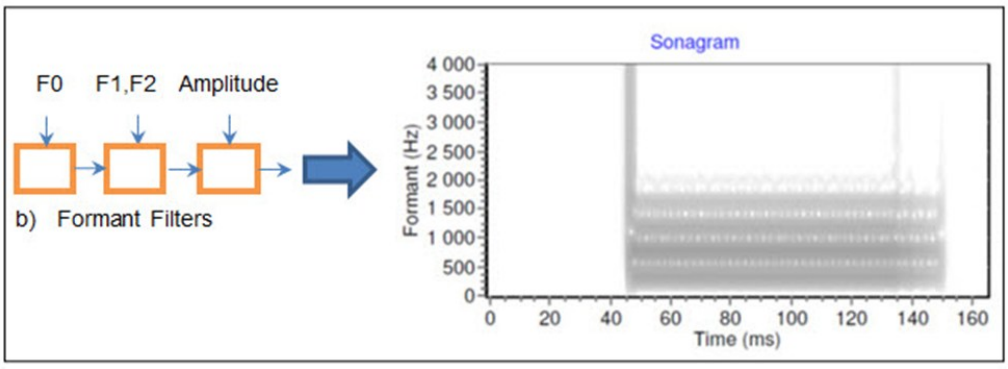


Figure 2-3: Formant synthesizer: the amplitude gate is placed after the formant circuits; and its sonogram of synthetic /a/, F0=400Hz.

- In these two experiments, we observe the spectrogram (in the right side of Figure 2-2 and 2-3) of the same vowel /a/ which is synthesized by the two first formant frequencies F1, F2 with the same F0 constant equal to 400 Hz.

2.2.1.2 Observation

In the first case (see in the left diagram of the Figure 2-2), the amplitude control is located before the formant circuit, thus the formant circuit responses are affected by the amplitude transients. Observing the spectrogram of the vowel /a/ in the right side of the Figure 2-2, we see that in the beginning and at the end of the amplitude transients, the spectrum of the output vowel /a/ is continuous and gives information on the whole transfer function. After that, the steady state condition corresponding to a harmonic structure spectrum is achieved. Because of the high fundamental frequency, the harmonic here is poor and the two first formants are not well defined except at the beginning and the end of the vowel /a/ due to the impulse response. In this case, a real simulation of the speech production is obtained.

In the second case (see in the left diagram of Figure 2-3), the amplitude gate is located after the formant circuits, that means the formant filters are permanently connected to the excitation source, therefore they are continuously excited by the pulse generator. The spectrogram of the vowel /a/ in the right side of the Figure 2-3 shows that the spectrum is composed of harmonics but there is no impulse response and the formant structure cannot be observed.

It is clearly seen on the two above figures that the impulse responses at the beginning and the end of spectrogram of vowel /a/ only appear with the front location of the amplitude circuit and they do not appear with the case of the after location's one. These observations agreed with the study of Carré and Lancia's study (1975); see also Carré and Quach (1987).

2.2.1.3 Conclusion

From the result obtained in experiment 1, and reminding that in speech production the amplitude control is placed before the vocal tract, we suggest that the impulse responses are produced in natural speech. Amplitude transients when they are effective (that means when they are produced before formant filters), and contains information about the whole spectrum (impulse responses). Therefore, we continue to suppose that any impulse response plays the same role, including in the transition part. We can propose to verify this supposition in the experiment 2.

2.2.2 Experiment 2 – Impulse response during the vocal tract transitions

2.2.2.1 Methodology

In order to study the role of impulse response during vocal tract transitions, we synthesize a speech sound V1V2 from the impulse response of the V1V2 transition by a combination system of two

formant synthesizers as in Figure 2-4. The two formant filter channels are used in this experiment having the different orders of formant filter: one with F2-F1 order and one with F1-F2 order, in which:

- The top formant filter channel (with the order of F2-F1): F1 filter is followed by F2 filter.
- The below formant filter channel (with the order of F1-F2): F2 filter is followed by F1 filter.

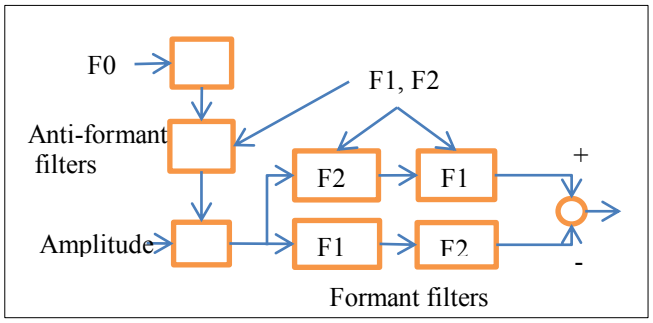


Figure 2-4: Formant synthesizer. Anti-formant circuits placed in the source of the synthesizer cancel the effects of the formant circuits during static functioning.

The synthesized V1V2 output is created from the difference between the output signal from the top formant channel and the output signal from the bellow formant channel.

In this experiment (Figure 2-4), anti-formant filters are placed between the pulse generator and the amplitude gate. Their anti-resonant frequencies are fixed at F1, F2 of the formant filters in order to get a flat spectrum at the output. F1 and/or F2 can be cancelled or not by corresponding to the antiformants in the source.

2.2.2.2 Results

When the top formant filter channel is used, F2 filter is followed by F1 filter. In this case, the output of F2 us non-stationary and leads to an impulse response of F1. Figure 2-5 gives an example of the spectrogram of the same preceding synthetic vowel /a/ as in the experiment 1, their impulse responses of the formant circuits are preserved. The formant structure disappeared during the stable part of the vowel at the advantage of harmonics corresponding to impulse spectrum. This means that whatever the shape of the vocal source, a more or less strong impulse response is obtained.

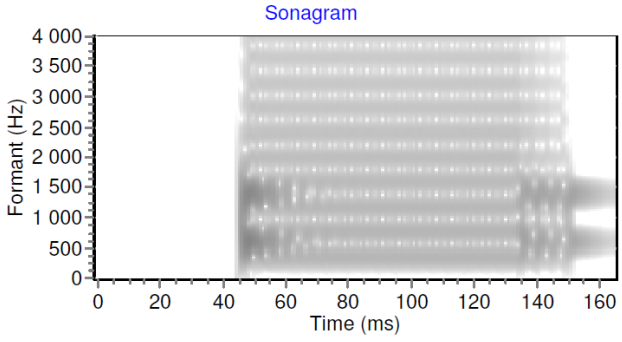


Figure 2-5: Sonogram of synthetic [a], F0=400Hz with an anti-formant filters placed in the source to get a flat spectrum. The impulse responses at the beginning and at the end of the signal can be observed.

Returning our experiment in Figure 2-4, we create a synthesized /ai/ with high F0, F0 = 400Hz. The spectrograms of synthesized speech sound of /ai/ from the top and bellow formant filter are observed in Figures 2-6 and 2-7, respectively. The obtained speech sounds of /ai/ in two cases hear clearly.

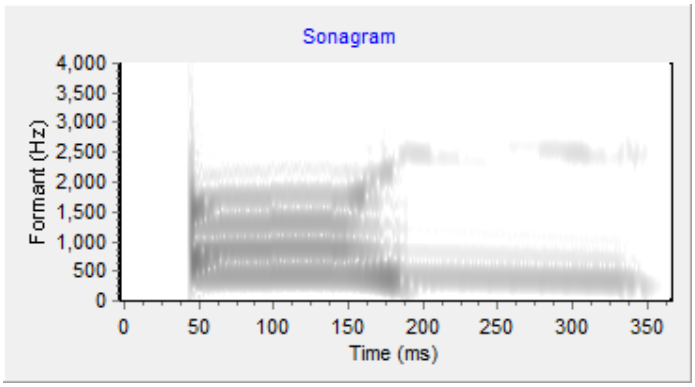


Figure 2-6: Sonogram of /ai/ from the formant filter of order F2-F1 (F0 = 400Hz).

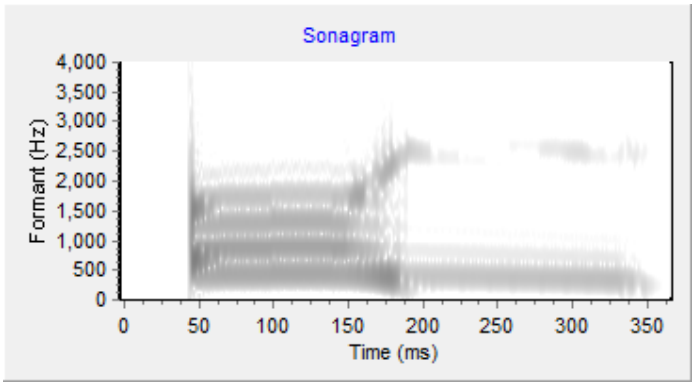


Figure 2-7: Sonogram of /ai/ from the formant filter of order F1-F2 (F0 = 400 Hz).

At the final output signal /ai/, the subtraction cancels the stable and the transitory parts of the signal but the impulse responses are still preserved during the transition parts. The spectrogram of this output signal is shown in Figure 2-8. Obviously, we can see clearly the spectrum of the impulse response due to the transition for /ai/.

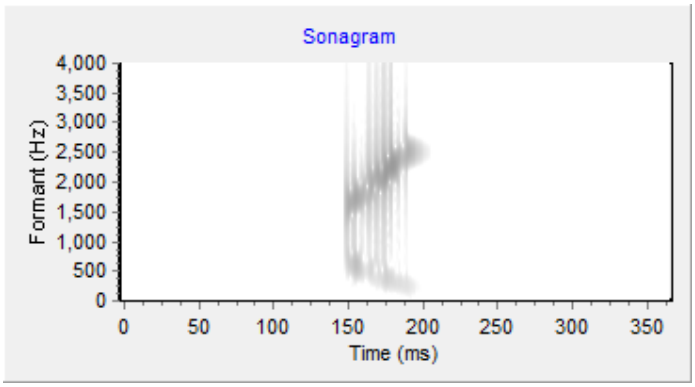


Figure 2-8: Sonogram of the impulse responses of the formant circuits due to the transition /ai/ after amplify the amplitude more 10dB.

This final speech signal /ai/ is clearly intelligible as /ai/ but the amplitude of the impulse transition is very weak, about 10dB less than the ones of each above formant filter channel.

2.2.2.3 Discussion

Observing the spectrogram of the synthesized /ai/ in figure 2-6 and 2-7, we see that these speech signals bring all information about the stable and transition parts of vowel /a/ and /i/, and these sound can be perceived clearly like /ai/. However, these spectrogram is not good, so it is difficult for us to extract useful information from these spectrum.

Moreover, the spectrogram of the final output signal /ai/ is obtained from only the transition part of /ai/. Although the amplitude of this impulse transition is very weak, but when we amplify it 10dB more, this transition spectrogram is fairly clearly, and the corresponding sound is still intelligible as /ai/. This is dynamic speech.

2.2.2.4 Conclusion

The result of experiment 2 provides evidence that speech signal can be intelligible from only impulse transition without stable information. However this information exists in the signal which is very small in amplitude (-10dB with the normal situation of formant filter) and, perhaps it is not detected. Therefore, we have to find other dynamic parameters, and phase spectrum is the next candidate. That is why we study the role of phase spectrum in speech perception in the experiments 3 and 4.

2.2.3 Experiment 3 – Speech signal characterization from power spectrum and phase spectrum

2.2.3.1 Methodology

Though speech is a non-stationary signal, it can be assumed to be quasi-stationary. Therefore, it can be processed through a short-time Fourier analysis. According to this approach way, the speech signal is completely characterized by its power spectrum and phase spectrum (Paliwal and Alsteris, 2003).

In order to verify this issue, we consider the module and phase of a transfer function of the vocal tract for the neutral position (i.e. F1=500Hz, F2=1500Hz, F3=2500Hz) in speech synthesis because in a formant synthesizer, amplitude and phase are correctly reproduced. It is well known that with such circuits, there is theoretically a strict correspondence between module and phase: at the resonant frequencies correspond phase transitions of -л and the stiffness of the phase transitions are proportional to the Q factors. Then we calculate the amplitude and phase spectrum of this signal. These two parameters are showed in Figures 2-9 and 2-10, respectively.

2.2.3.2 Observations and discussions

Simultaneous combination of two Figures 2-9 and 2-10 reveals that the phase value corresponding to each formant frequency is different. This result agrees with the study of Paliwal and Alsteris's study (2003). And it means that phase information is theoretically sufficient to recover the transfer function. For example, Figure 2-11 shows the spectrum of a synthesized speech /ai/ obtained from harmonic components with equal amplitude but with phase variations. Indeed, traces of formant variations can be observed. Remember that phase variations theoretically lead to frequency variations. Moreover, the formant frequency variations can be observed in the sonogram, but they cannot be perceived due to the masking effect of the flat general spectrum (equal component amplitudes). This point must be further studied.

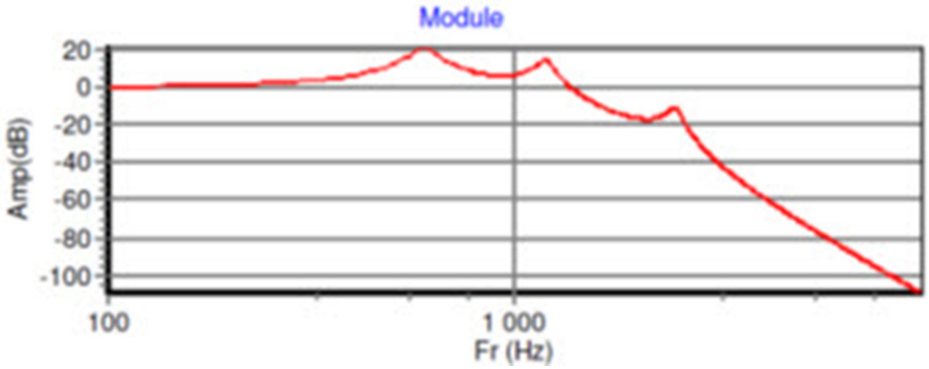


Figure 2-9: Module of the transfer function for a neutral position.

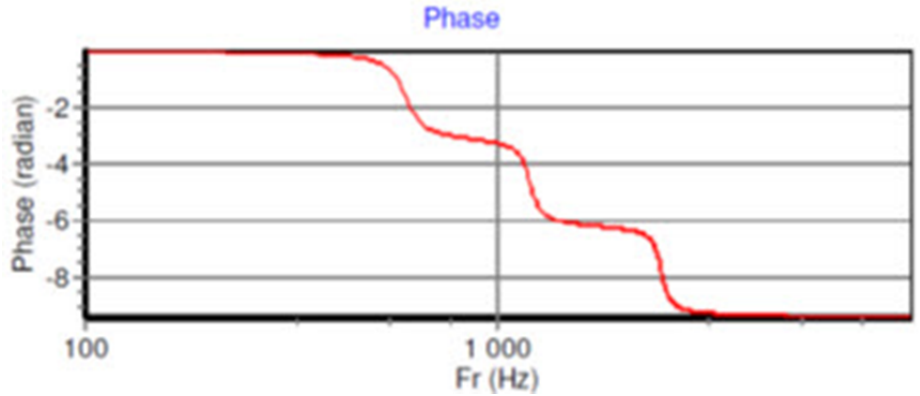


Figure 2-10: Phase of the transfer function for a neutral position.

Moreover, the studies of Yegnanarayana and Murthy (1992) and Alsteris and Paliwal (2007) showed that it is difficult to determine phase spectrum in natural speech.

According to these above discussions, a key question raises: what is the role of the phase variations in speech perception? Is the amplitude spectrum of speech signal, without the information of phase spectrum, sufficient to recognize speech sound? Experiments 4 presented in section 2.2.4 will answer this question.

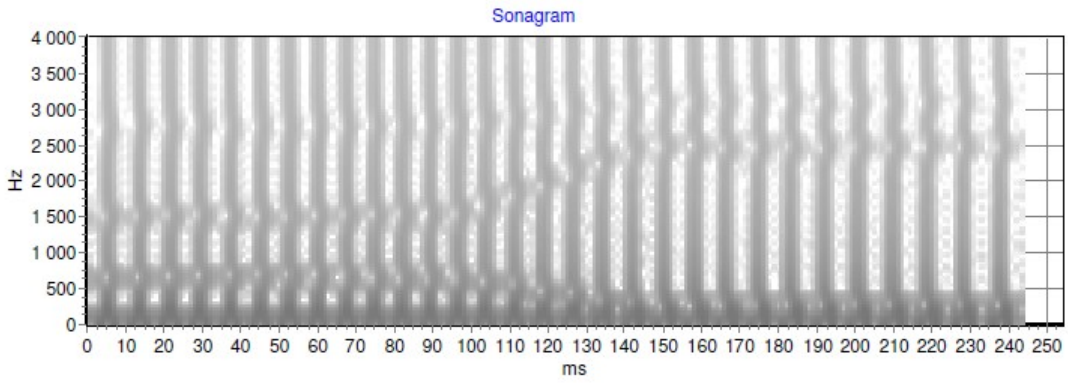


Figure 2-11: Sonogram of synthetic /ai/. *Harmonic component amplitudes are fixed. The formant phases vary according to the scheme shown in Figure 2-10. Formant variation traces are observed.*

2.2.4 Experiment 4 – The role of amplitude spectrum in perceptive speech

Preliminary perception tests on experiment 1 and 2 showed that the effects of the impulse responses (at the beginning, at the end and during the transition) are slight. The phase variation can be important according to Paliwal and Alsteris's study (2003). In our experiment 3, formant frequency variations can be observed in the sonogram (Figure 2-11) but they cannot be perceived due to masking effect of the flat general spectrum (with equal amplitude). Therefore, this point must be further studied in the following content.

2.2.4.1 Objective

The aim of this experiment is to evaluate the speech perception of synthesized speech /bVk/ if we only use the amplitude information and without information of phase (in which V is one of three Vietnamese short vowel /ɤ̆, ɔ̆, ă/).

2.2.4.2 Methodology

2.2.4.2.1 Stimuli

In order to implement this experiment, we perform the perception tests using the three Vietnamese words: /bɤ̆k/, /bɔ̆k/ and /băk/, created by a female voice in the three following modalities:

- First modality, these speech signals are pronounced by a North-Vietnamese female native speaker.
- Second modality, we synthesize three speech signals by formant synthesis. This synthesis method implementation is the same acoustic production process as in natural speech.
- Third modality, these signals are synthesized with harmonic components (computed according to experiment 3 of section 2.2.3). In this case, the phases between harmonic components are fixed and equal to 0.

For the second and third synthesizers, the variations of the three first formant frequencies are obtained from the natural speech signals of the first synthesizer. For all modalities, F0 is high (around 300 Hz, i.e. female voice).

2.2.4.2.2 Perception test

Six Vietnamese participants (3 males + 3 females) took part in this perceptual test. These subjects are from the north dialect with the ages from 25 to 35 years old.

Each item in each modality is presented ten times in random order.

For each additive stimulus, each participant had to give a choice among four possibilities:

- One of the three short vowels /ɤ̆/, /ɔ̆/ or /ă/.
- The final choice is no acknowledgment when the participant could not identify any vowels among three short vowels /ɤ̆, ɔ̆, ă/.

2.2.4.3 Results and discussions

The perceptual results will be computed in percent (%). Figure 2-12 shows the perceptual result of three short Vietnamese vowel pronounced by a Vietnamese female native speaker. These natural items are clearly recognized for all vowels (100%). The items obtained by formant synthesis of the second modality are also clearly recognized (close to 100%, see Figure 2-13). In the third modality (Figure 2-14), the speech synthesizer is performed by changing only the amplitudes of the harmonic components without information of its phases.

The obtained result in the third modality shows that the three Vietnamese short vowels are still clearly perceived when the information of phase variations in the vowel speech synthesis (the phases between harmonic components are fixed and equal to 0) is ignored. This result suggests that the amplitude of the spectral components is clearly sufficient for perceptual discrimination of vocalic sounds.

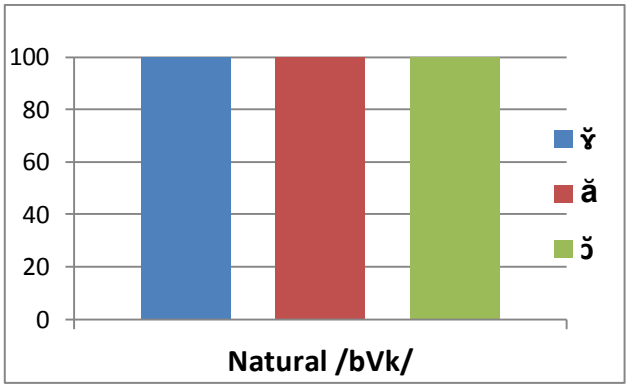


Figure 2-12: Identification score for the three Vietnamese short vowels pronounced by a Vietnamese female native speaker (%).

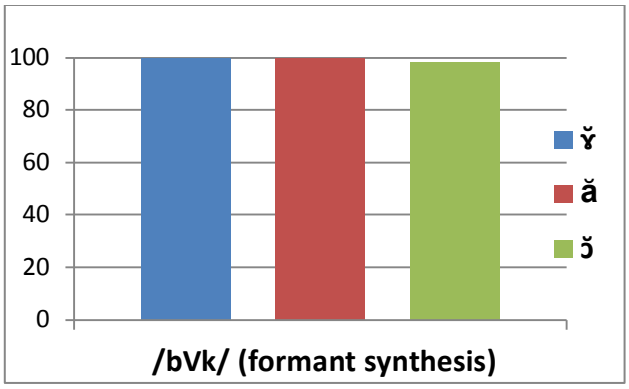


Figure 2-13: Identification score for the three short Vietnamese vowels obtained from a formant synthesizer with F0 around 300Hz (%).

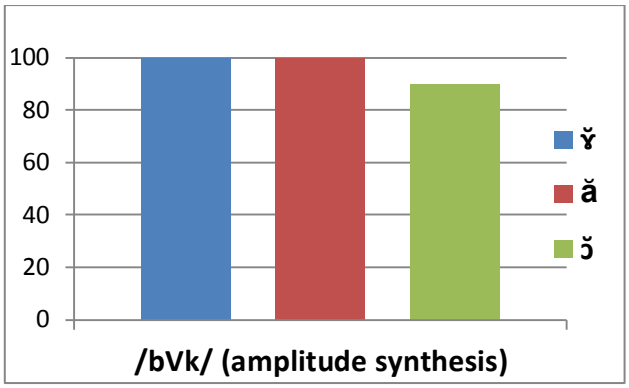


Figure 2-14: Identification score for the three Vietnamese short vowels obtained from a synthesizer with only the amplitude of harmonic components and without their phase information (%).

2.3 Conclusions of chapter 2

Chapter 2 studied some speech characteristics on both static and dynamic aspects both in the speech signal and in the perception test. Both natural and synthesized speech signals were obtained with high fundamental frequency (i.e. female voice) whose harmonic spectrum is poor and the formant frequency cannot be well detected.

In experiments 1 and 2, we studied the effects of the impulse response in the beginning, at the end and during transition of the speech signals. The obtained results showed that the impulse responses are produced in natural speech and their effects of these parameters can be observed in spectrograms. However, this information exists in the signal which is very small in amplitude (-10dB with the normal situation of formant filter) and, perhaps it is not detected. Therefore, we continued to study the effects of phase spectrum (which is considered next dynamic characteristic candidate of speech) in perception test in experiment 3 and 4.

Experiment 3 studied on synthesized speech signal in order to show that the speech signal is completely characterized by its power spectrum and phase spectrum. This agreed with the study of Paliwal and Alsteris (2003). However, in fact, it is difficult to extract phase spectrum in the natural

speech signals. Experiment 4 was performed in order to solve the question on the role of the phase variations in speech perception. This experiment is based on a perception test of the three Vietnamese short vowel (/ɤ̆,ɔ̆,ă/) in the synthesized speech signals of /bɤ̆k/, /bɔ̆k/ and /băk/ if the amplitudes of the harmonic components are the main parameter and the phases between harmonic components are fixed and equal to 0 or without the information of phase spectrum. The obtained results showed that in case of high F0 (i.e. female voice), the formant frequencies for the three Vietnamese short vowels were not detected from amplitude spectrum but synthesized speech signals obtained from amplitude spectrum (without phase information) were clearly perceived showing that the spectral amplitude information can be sufficient for perceptive speech discrimination without phase information. This means that formant frequencies were not detected by the human perception system. With regard to phase, Paliwal and Alsteris (2003) showed that syntheses with either spectral amplitude only or spectral phase information only lead to signals clearly recognized.

Keeping up the results obtained in Chapter 2, the work of next chapters is finding out another dynamic candidate of speech signal (not phase spectrum) that can be useful for speech discrimination. And these parameters will be extracted only on amplitude domain and without taking into account the phase information of speech.

Chapter 3
Dynamic acoustic characteristics at the speech perception level

In phonology, a phoneme is the smallest discrete (or distinctive) unit of speech that distinguishes meaning and that one can isolate by segmentation in the spoken language. A phoneme is an abstract entity, which can correspond to several sounds, referred to as the phoneme's allophones. Each language has a particular set of phonemes; importantly, each phoneme's range of allophonic variation, too, is language-specific. In general, this set of phonemes can be classified into small groups according to their articulatory and/or acoustic characteristics.

A vowel is a type of sound for which there is no constriction or closure at the throat or mouth at any point during the interval where vocalization occurs. Vowels can be contrasted with consonants, which are characterized by an obstruction (occlusion or constriction) at one or more points along the vocal tract.

In the acoustic domain, vowels are generally characterized by the first two or three formant frequencies. Each of them can be represented as a dot in the acoustic space (F1-F2 and F2-F3 planes) (Peterson and Barney, 1952; Vallée, 1994; Gendrot and Adda-Decker, 2007; Neel, 2008; Vaissière, 2011) and specified in terms of underlying 'targets': context- and duration-independent formant values as obtained by fitting "decaying exponentials" to the data points (Moon and Lindblom, 1994). The point in focus here is that this specification is static and may be taken to imply that the perceptual representation corresponds to the target values (Strange et al., 1983; Strange, 1989b). But the vowel targets depend on the speaker, and are rarely reached in spontaneous speech production (Lindblom, 1963a; Carré, 2008; Website 3, 2009). Moreover, vowel's acoustic characteristics are different according to the context and the speaker (Nordström and Lindblom, 1975; Verbrugge and Rakerd, 1980; Strange, 1989b; Johnson, 1990; Johnson et al., 1993; Johnson, 1997; Carré and Hombert, 2002). Therefore, how is the vowel perceived with its different acoustic characteristics according to the

context and the speaker? Numerous studies have been devoted to this question (Nordström and Lindblom, 1975; Strange, 1989b; Johnson, 1990). The results remain somewhat incomplete and contradictory. However they contribute to highlight the importance of the dynamics in vowel perception (Shankweiler et al., 1978; Verbrugge and Rakerd, 1980; Strange et al., 1983). Among these studies, that one conducted by Carré and colleagues (1991) is presented in details in section 1.4.2. It deals with the characterization of speech dynamics by the direction and rate of vocalic transitions (René Carré and Mrayati, 1991). Carré and Mrayayi showed the role of speech dynamics on vowel-to-vowel (V1V2) perception. They focused on the direction and rate of synthesized transitions situated outside the F1/F2 vowel triangle defined in absolute values. This situation enables the study of transitions characterized only by their directions and rates without reference to any vowel targets in the vowel triangle. The results showed that the perception of V1V2 is based on the direction and length (i.e. rate of the transition) of the trajectories. This study and others provide evidence on the important role of transitions in the identification of vowel qualities.

However one question here is what happens to this result if vowel-consonant-vowel sequences are considered? To what extent will the same findings be replicated when stimuli have an intervocalic consonant? Chapter 3 proposes to use the same kind of synthetic experiment leading to the perception of vocalic acoustic both non-illusions and illusions for the study of consonant perception.

To summarize the results of experiments presented in Chapter 2, the amplitudes of the spectral components seem important and clearly sufficient for the perceptual discrimination of VV sequences without phase information. In view of this finding, in Chapter 3, we focus on the amplitude domain without taking into account the phase information of speech.

3.1 Introduction

As mentioned above, vowels are generally described with static articulatory configurations represented by targets in the acoustic space: typically, as a dot in formant frequencies plane of the F1-F2 and F2-F3 planes (Peterson and Barney, 1952; Vallée, 1994; Gendrot and Adda-Decker, 2007; Neel, 2008; Vaissière, 2011). Plosive consonants can be described in terms of places of articulation, represented by locus or locus equations in an acoustic plane (Lindblom, 1963b; Sussman et al., 1991; Sussman and Shore, 1996; Brancazio and Fowler, 1998; Graetzer, 2008; Nguyen et al., 2008; Nguyen, 2009).

However, formant frequencies for vowels vary considerably with consonantal context (co-articulation) and with speaking rate/reduction phenomena (Lindblom, 1963a; Agwuele et al., 2008; Lindblom and Sussman, 2012). Several studies have been proposed to solve this problem by speaker normalization (Nordström and Lindblom, 1975; Johnson, 1990; Johnson, 1997; Strange, 1989b). Speaker normalization techniques modify the spectral representation of incoming speech waveforms in

an attempt to reduce variability between speakers. The use of speaker-specific feature warping and frequency warping are typical methods. However choosing a suitable warping function (or factor) for each speaker is difficult and this work is generally incomplete.

Moreover, sensory systems have been shown experimentally to be more sensitive to changing stimulus patterns than to purely steady-state one (Pollack, 1968; Divenyi, 2005). In this light, it appears justified to look for an alternative to static targets: a specification that recognizes the true significance of the variation of the signal over time. One possibility is that dynamics can be characterized by the direction and the rate of vocalic transitions. Vowel-vowel trajectories in the F1-F2 plane are generally rectilinear (René Carré and Mrayati, 1991). Therefore, they can be characterized by their directions.

On the topic of transition duration, we recall the results of Kent and Moll (1969): "*the duration of a transition – and not its velocity – tends to be an invariant characteristic of VC and CV combinations*". Gay (1978) confirmed these observations with different speaking rates and with vowel reduction, namely the reduction in duration during fast speech is reflected primarily in the duration of the vowel, and the transition durations within each rate were relatively stable across different vowels. These same points were observed in the Vietnamese structures VC and CV (Nguyen, 2009). If the transition duration is invariant across a set of CVs with a constant C and varying Vs, the transition rate depends on the vowel to be produced. At the very beginning of the transition and throughout the transition there is sufficient information to detect the vowel to be produced. If the perception of the following sound is based on the syllabic duration, on the transition direction and rate, then we can explain the perceptual results obtained by Strange (1983) in 'silent center' experiments that replaced the center of the vowel by silence of equivalent duration. This manipulation preserves the direction and the rate of the transition as well as the temporal organization (syllabic rate).

In a perceptual study, Carré (2009a) showed that the synthesized transitions situated outside the traditional F1-F2 vowel triangle were perceived as vowel-to-vowel transitions. In this case, there is no reference to any vowel targets in the vowel triangle. The results of this study can be summarized by saying that the region where 4 trajectories converge (acoustically closed to /a/, in absolute frequencies) was perceived as /a/ or /u/ or /o/ depending on the direction and length (i.e. rate of the transition) of the trajectories. In order to extend this result, we intend to use the same kind of synthetic experiment leading to the acoustic perception of consonant in both non-illusion and illusion cases in the following sections.

3.2 Consonant perception in pseudo-V1CV2

3.2.1 General methodology

3.2.1.1 Type of experiments

In order to study consonant perception, two experiments (non-illusion test and illusion test) are realized in which a V1CV2 item is synthesized by formant frequencies. The main difference between these two tests is the acoustic context of the consonant C, that is, vowels V1 and V2:

- for the non-illusion test, vowel V1 and V2 are placed inside the vowel triangle.
- for the illusion test, these vowels are located outside the vowel triangle.

3.2.1.2 Perceptual test process

Both experiments (non-illusion and illusion) are carried out with ten Vietnamese native participants (5 men: M1, M2, M3, M4, M5 and 5 women: W1, W2, W3, W4, W5) from the north of Vietnam, their age from 25 to 35 years old. They speak Vietnamese daily. None of them had training at university in phonetics-phonology. None was aware of the purpose of the two experiments.

The subjects were instructed to listen to the V1CV2 stimuli with headphones, then identify the perceived consonants by clicking the response on the screen using the mouse in the corresponding program.

Each participant had to listen to three times 60 V1CV2 stimuli presented in random order, and chose which consonant they perceived.

Four possibilities of responses were given for perceptual results, as follows:

- Three consonants, among: b, d and g. These three letters correspond, in the Vietnamese alphabet, to the phonemes /ɓ/, /ɗ/ and /ɣ/, respectively: two stop consonants and a fricative one, whose point of articulation is labial, dental and velar, respectively.
- The fourth choice was the choice “NAK” (non-acknowledgment), selected when the listener does not perceive any of /ɓ/, /ɗ/ or /ɣ/ in the stimulus.

3.2.2 Non-illusion experiment

3.2.2.1 Purpose

The non-illusion experiment was performed to find out the region of the consonants /ɓ, ɗ, ɣ/ in the F1-F2 and F2-F3 planes in synthesized V1CV2 context in which the vowels V1, V2 are situated inside the vocalic triangle.

3.2.2.2 Stimuli

In this experiment, there are 60 sounds of V1CV2 items that were synthesized by the third first formant frequencies, in which the two vowels V1 and V2 are very close to vowel [i] and vowel [a] inside the vowel triangle, respectively.

V1 and V2 durations were 100ms, and 120ms, respectively. The durations of the transition V1C and CV2 were 30ms.

The consonant C is synthesized without burst for different formant values: F1 varies by equal step of 100Hz between 100Hz and 300Hz; F2 varies by equal steps of 500Hz between 500Hz and 2500Hz; and F3 varies by equal steps of 500Hz between 2000Hz and 3500Hz.

3.2.2.3 Results

Ten Vietnamese subjects, as described in detail in section 3.2.1.2, participated in this experiment. The perception test process was performed as in section 3.2.1.2. Table 3-1 shows the main results of the perception test. The average scores of correct identification were calculated for the ten participants.

Firstly, the score of NAK is very small. This weak score observed for NAK responses suggests that all the participants can easily identify the three consonant /ɓ, ɗ, ɣ/.

Secondly, for these three consonants, a high score of identification is observed: /ɓ/, /ɗ/, and /ɣ/ are 100%, 94%, and 93%, respectively; /ɓ/ is perceived more correctly than the others /ɗ/ and /ɣ/ (with the highest score, and no confusion with both /ɗ/ and /ɣ/); the consonant phonemes /ɗ/ and /ɣ/ are less distinct with some confusions; the region of the consonant phoneme /ɗ/ is smaller than the one of the consonant phonemes /ɓ/ and /ɣ/;

In spite of some confusions between the consonant phoneme /ɗ/ and /ɣ/, we still distinguish the three regions corresponding to these three consonants.

Thirdly, the F2 formant plays an important role that makes possible the discrimination of the three consonants /ɓ, ɗ, ɣ/; this is in agreement with the results obtained in previous studies of Liberman (1954), Serniclaes (1987) and Nguyen (2010).

Figure 3-1 shows a representative example of the perceptual results of two participants in the F1-F2-F3 space: one is the result of female W1 and one is the result of male M1. In this figure, the blue round corresponds to the consonant /ɓ/, the red cross corresponds to the consonant /ɗ/ and the green square corresponds to the consonant /ɣ/. A sign (circle, cross or square) will be marked in the 3-D space if the correct recognition rate of the consonant (/ɓ/, /ɗ/, or /ɣ/) is higher than 50%. The sign dimension is also proportional to the correct recognition rate value.

Table 3-1: Main identification scores in the non-illusion experiment. The average correct recognition rates are calculated for ten subjects

Formant			Correct recognition rate			
F1	F2	F3	/ɓ/	/ɗ/	/ɣ/	NAK
100	500	2000	100%	0%	0%	0%
200	500	2000	100%	0%	0%	0%
300	500	2000	100%	0%	0%	0%
100	500	2500	100%	0%	0%	0%
200	500	2500	100%	0%	0%	0%
300	500	2500	100%	0%	0%	0%
100	500	3000	100%	0%	0%	0%
200	500	3000	100%	0%	0%	0%
300	500	3000	100%	0%	0%	0%
100	500	3500	100%	0%	0%	0%
200	500	3500	100%	0%	0%	0%
300	500	3500	100%	0%	0%	0%
100	1000	2000	100%	0%	0%	0%
200	1000	2000	100%	0%	0%	0%
300	1000	2000	100%	0%	0%	0%
100	1000	2500	100%	0%	0%	0%
200	1000	2500	100%	0%	0%	0%
300	1000	2500	100%	0%	0%	0%
100	1000	3000	100%	0%	0%	0%
200	1000	3000	100%	0%	0%	0%
300	1000	3000	100%	0%	0%	0%
100	1000	3500	100%	0%	0%	0%
200	1000	3500	100%	0%	0%	0%
300	1000	3500	100%	0%	0%	0%
100	1500	2000	10%	84%	3%	3%
100	1500	2500	17%	80%	3%	0%
300	1500	2500	7%	87%	7%	0%
200	1500	3000	0%	89%	11%	0%
300	1500	3000	0%	94%	6%	0%
100	1500	3500	3%	91%	6%	0%
200	1500	3500	3%	70%	27%	0%
300	1500	3500	3%	85%	12%	0%
200	2000	2000	0%	16%	84%	0%
300	2000	2000	0%	15%	85%	0%
300	2000	2500	0%	33%	70%	0%
100	2500	2000	6%	23%	71%	0%
200	2500	2000	0%	10%	90%	0%
300	2500	2000	3%	3%	93%	0%
200	2500	2500	0%	17%	83%	0%
300	2500	2500	0%	18%	82%	0%
200	2500	3000	3%	12%	84%	0%
300	2500	3000	0%	25%	75%	0%
200	2500	3500	3%	16%	81%	0%
300	2500	3500	0%	18%	82%	0%

Observing Figure 3-1, we can see that both two participants can identify easily the three consonants /ɓ, ɗ, ɣ/ (the dimension of rounds, crosses and squares are great). The result on Figure 3-1 shows also that both W1 and M1 discriminate the three consonants by more or less the same values of F1, F2, and F3 (though there is a small overlap between the consonant /ɗ/ with /ɓ/ and /ɣ/ in the perceptual scores of M1, the three regions corresponding to the three consonants /ɓ, ɗ, ɣ/ are still distinct).

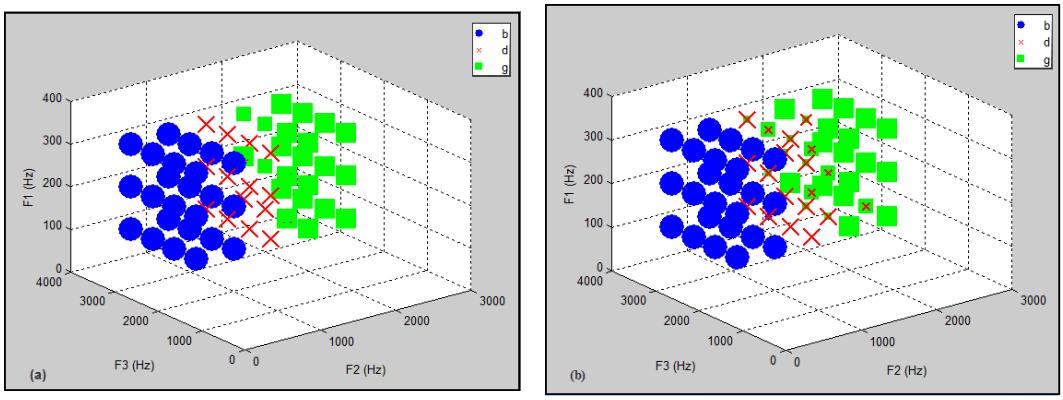


Figure 3-1: The perceptive results of two subjects in the non-illusion experiment: one female W1 in (a), and one male M1 in (b).

3.2.3 Illusion experiment

3.2.3.1 Purpose

The illusion experiment was performed to find out the region of the consonants /ɓ, ɗ, ɣ/ in the F1-F2 and F2-F3 planes in V1CV2 context in which the pseudo-vowels V1, V2 are situated outside the vocalic triangle.

3.2.3.2 Stimuli

There are also 60 sounds of the V1CV2 items were synthesized by the third first formant frequencies, in which the trajectory of the sequence V1V2 were fairly parallel to the trajectory [i-a] in the vowel triangle: the F1, F2 values of pseudo-vowel V1 were 1000Hz, 1490 Hz, respectively, and the ones of pseudo-vowel V2 were 420Hz, 2680Hz, respectively. The durations of V1 and V2 were 100ms, and 120ms, respectively. The durations of V1C and CV2 were 30ms. The consonant C was synthesized without burst for different formant values yielding 60 stimuli of V1CV2 items: F1 varies by equal step of 150Hz between 300Hz and 600Hz; F2 varies by equal step of 1500Hz between 1000Hz and 7000Hz; and F3 varies by equal step of 1500Hz between 4000Hz and 8500Hz.

Comparing with the formant values in non-illusion test, although the absolute values of F1 and F2 of V1 or V2 is different in two experiments, the transition direction and rate of V1V2 is more or less the same in both experiment. And the formant grids of consonant C in case of illusion test are the shift with fast transition of the one in non-illusion test in both F1-F2 and F2-F3 plane. This point is expressed by the bigger changing step of each formant in the illusion test.

Figure 3-2 shows the corresponding above location of two vowels [V1] and [V2] outside the vocalic space (that was expressed by green triangle), and the consonants were expressed by the round points. The vowel points are below the blue line in F1-F2 plane having their F2 values are smaller than F1 ones; and the vowel points are below the red line in F2-F3 plane having their F3 values are smaller

than F2 ones. All those vowels are not suitable with vowel theory (F1 value < F2 value < F3 value), therefore, those vowels will not be studied in our experiment.

This experiment was carried out with the same ten Vietnamese participants as in the description of section 3.2.1.2, and with the same choices for the perceptual test as in the non-illusion experiment (previous section: 3.2.2).

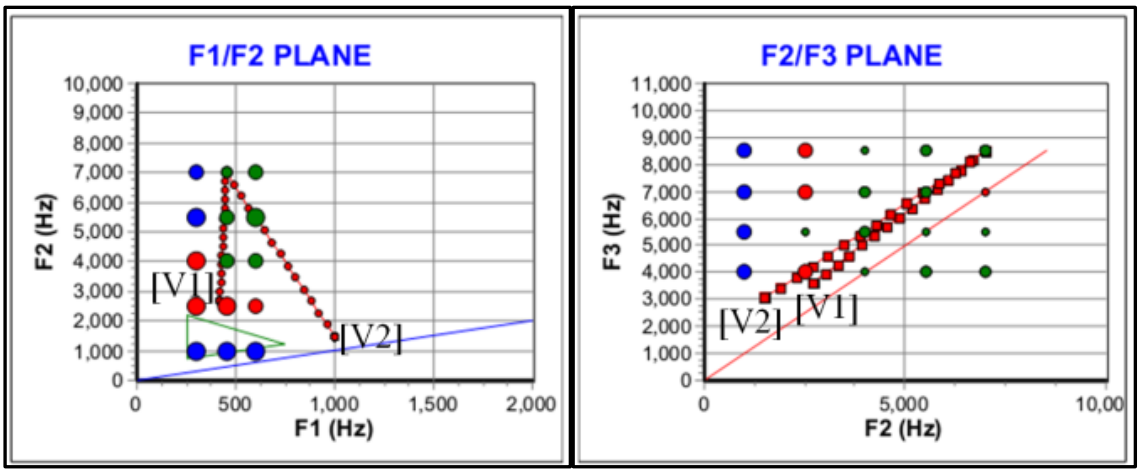


Figure 3-2: The V1CV2 stimuli in the illusion experiment: the pseudo-vowels V1, V2 are situated outside the vocalic space (green triangle); the consonant C is synthesized without burst for different formant values of F1, F2, F3.

3.2.3.3 Results

Most participants still correctly identified the three consonants /ɓ, ɗ, ɣ/. However, one participant (M5) could not recognize the consonants /ɗ, ɣ/: his responses are either /ɓ/ or NAK. The results of M5 called for a separate analysis. We decided to calculate the average correct recognition rates of nine subjects (without M5). The main results are shown in Table 3-2.

Firstly, most participants still recognize the three consonants /ɓ, ɗ, ɣ/, but with a worse score than the one in the non-illusion test because the perceptual test in this case are performed in an abnormal situation

Secondly, the region of the consonant /ɗ/ is still smaller than the one of the consonants /ɓ/ and /ɣ/. This result points in the same direction as the result in non-illusion test in section 3.2.2.

Thirdly, in spite of some confusions among the three consonants /ɓ/, /ɗ/, and /ɣ/, we still identify three distinct regions corresponding to these three consonants.

Finally, although this test was carried out in an abnormal context where the pseudo-vowels V1, V2 are placed outside the vocalic triangle, F2 plays again an important role to discriminate the three consonants /ɓ, ɗ, ɣ/. This result is keeping with that obtained in a normal context of non-illusion test.

As in non-illusion test, in order to provide a bird's-eye view of the result, Figure 3-3 shows a representative example of the perceptual results of two participants in the 3-D space of F1-F2-F3: one

is the result of female W1 and one is the result of male M1. Similar to the Figure 3-1 of non-illusion test, the blue round, the red cross and the green square are corresponding to the consonant /ɓ/, /ɗ/, / ɣ/, respectively. And the sign dimension is also proportional to the correct identification rate value.

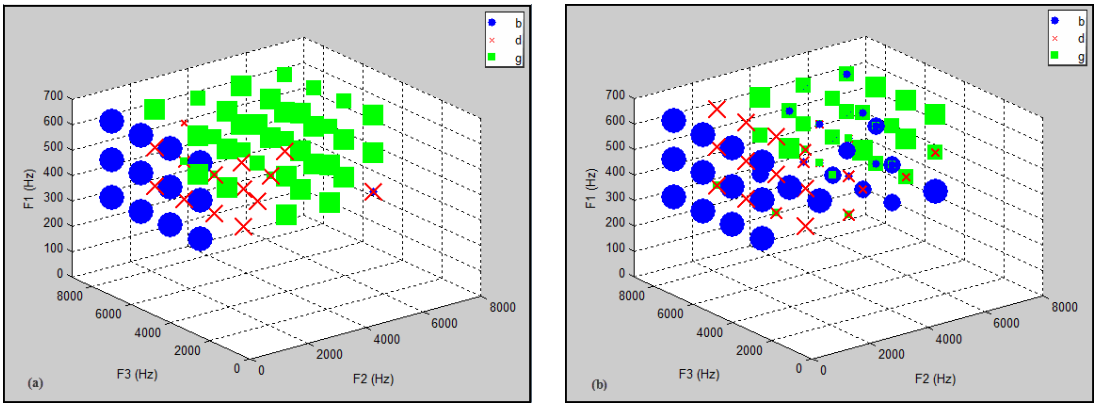


Figure 3-3: The perceptive results of two subjects in the illusion experiment: one female subject W1 in (a), and one male M1 in (b).

Table 3-2: Main identification scores in the illusion experiment. The average correct recognition rates are calculated for nine subjects (without the result of the participant M5).

Formant			Correct recognition rate			
F1	F2	F3	/ɓ/	/ɗ/	/ɣ/	NAK
300	1000	4000	85%	13%	3%	0%
450	1000	4000	93%	4%	3%	0%
600	1000	4000	78%	0%	6%	17%
300	1000	5500	92%	8%	0%	0%
450	1000	5500	97%	3%	0%	0%
600	1000	5500	74%	4%	6%	17%
300	1000	7000	92%	8%	0%	0%
450	1000	7000	97%	0%	3%	0%
300	1000	8500	92%	8%	0%	0%
450	1000	8500	96%	0%	4%	0%
300	2500	4000	19%	74%	10%	0%
450	2500	4000	0%	88%	15%	0%
300	2500	5500	11%	71%	18%	0%
450	2500	5500	0%	92%	8%	0%
450	2500	7000	0%	93%	7%	0%
450	2500	8500	6%	92%	3%	0%
600	4000	4000	4%	10%	86%	0%
600	4000	5500	7%	13%	75%	8%
600	4000	7000	3%	18%	79%	0%
600	4000	8500	4%	6%	82%	8%
600	5500	4000	14%	19%	72%	0%
600	5500	5500	8%	7%	85%	0%
600	5500	7000	13%	7%	81%	0%
600	5500	8500	7%	4%	89%	0%
600	7000	4000	4%	7%	81%	8%
600	7000	5500	10%	19%	74%	0%
600	7000	7000	18%	4%	81%	0%
600	7000	8500	6%	15%	82%	0%

Visually, we see that the dimension of almost rounds, crosses and squares are great. Therefore, most of V1CV2 positions can be identified easily for both participants W1 and M1.

Although there exists some confusions between /ɗ/ with /ɓ/ and /ɣ/, but these confusions here are small: in particular, for the male subject M1, there is a small confusion between the consonant /ɓ/ with the consonant /ɣ/ where the consonant /ɓ/ is identified with very high values of F2 and small ones of F3.

3.3 Discussion

The dynamic approach is attractive because it potentially allows for the integration of consonants and vowels within a single theory. Conceivably, using the parameter of transition rate, one might propose that fast transitions tend to produce consonants, whereas slow transitions produce vowels.

In the case of perceiving V1V2 sequences, acoustic measurements indicate that signal information on V2 is available throughout the transition and especially at its very beginning. This strategy presupposes that the identity of the previous V1 has been determined.

In this study, the results of both experiments (non-illusion and illusion) show that the Vietnamese participants can identify and discriminate the three consonants /ɓ, ɗ, ɣ/ in the two different vowel contexts (inside and outside the vocalic triangle). In the illusion experiment, the three consonants /ɓ, ɗ, ɣ/ are perceived with more difficulty (because of the smaller correct recognition rates) than in the non-illusion experiment. This can be explained by the fact that the perceptual tests in the illusion experiment are carried out in an abnormal situation.

However, both experiments emphasize the important role of F2 in distinguishing the three consonants /ɓ, ɗ, ɣ/.

Comparing the results of the two experiments, we can realize that as F2 increases (from 500Hz to 2500Hz in the non-illusion experiment, or from 1000Hz to 7000Hz in the illusion one), the order of the three distinct regions corresponding to the three consonants /ɓ, ɗ, ɣ/ does not change: the consonant /ɓ/ is perceived with the lowest F2 values, and the consonant /ɣ/ is recognized with the highest ones.

Though there is an overlap between the F2 values in the non-illusion experiment and the illusion one (from 1000Hz to 2500Hz), the three consonants /ɓ, ɗ, ɣ/ are identified in succession in the non-illusion test, whereas only two consonants /ɓ, ɗ/ are identified at this same region of F2 in the illusion test. This confirms that the aspect of static F2 value is not important, but the relative F2 variation in relation to the pseudo-vowels and the dynamic value (i.e. rate of the F2 transition) of the trajectories V1C and/or CV2 play an important role.

The aim of the present study could suggest that dynamic parameters such as the transition direction and rate could be more invariant across males, females and children than vowel and consonant targets. When this hypothesis will be confirmed, it is unnecessary to make normalization in terms of static targets. However, normalization of transition rate with respect to the different transition durations observed in production (and depending on the speaker) would seem necessary. Such normalization could be readily available perceptually, thanks to temporal coding and the sensitivity of the auditory system to rate (derivatives) and acceleration (Pollack, 1968; Divenyi, 2005).

Our preliminary results on consonant perception with transitions outside the vowel triangle represent another step (after the perception of vowel-vowel transitions studied by Carré (2009a)) in support of a full dynamic approach.

Perception tests of formant transitions outside the vowel triangle encourage us to study general dynamic properties of the auditory system that may be used in speech.

3.4 Conclusions of chapter 3

In Chapter 3, we performed two consonant perception test in both non-illusion and illusion cases. The preliminary results presented here on consonants follow up the ones on vowels obtained by Carré (2009a) and clearly show the importance of dynamic characteristics for speech perception – which does not mean that static targets are not used in perception.

However, the dynamic approach needs to develop new ways of thinking and new tools. Formant transitions cannot be obtained from a succession of static values but from directions and slopes. It means that a new tool to measure directly these characteristics has to be developed. The dynamic approach is not a static approach plus dynamic parameters taken into account, it must be an intrinsically dynamic approach. With such an approach, in syllabic co-production, traditional static targets are extrinsic values whereas transition parameters become intrinsic values.

Chapter 4
Modeling dynamic acoustic speech features

As reviewed in Chapter 1, natural speech is not a simple succession of steady-state segments, but rather a dynamic process (Divenyi et al., 2006; Strange et al., 1983; Strange, 1989b). This holds true in speech production as well as in speech perception. And from there, some researchers started thinking to incorporate the dynamic aspects of speech into real applications. Some significant results were obtained, e.g. using spectral dynamics to improve the accuracy of speech recognition system (Hermansky, 2011). However, these dynamic concepts remain chiefly seen as ways to supplement static parameters: dynamic concepts are seen as deriving from static ones. And the function devoted to them is essentially that of reinforcing or accelerating processing methods that remain based on static parameters.

In Chapter 2, we did a short study of speech signal in terms of amplitude and phase. Our obtained results showed that the impulse responses (in the beginning, at the end and during the transition) are produced in natural speech, but their effects are slight.

Phase spectrum was studied in Chapter 2, as a first candidate of dynamic speech feature that is easy to capture in natural speech. However, the perceptual results showed that the amplitude information can be sufficient for perceptive speech discrimination without phase information. Therefore, we should try to find out another dynamic candidate of speech signal (not phase spectrum) that can be useful for speech discrimination.

Keeping up the role of only amplitude information in perceptual speech, we only analyze speech characteristics on amplitude domain, leaving out the phase information of speech in our following studies.

As reported in Chapter 3, we ourselves performed more speech perception tests on consonants. The present chapter extends the investigation to consonants perception here: two experiments were carried

out to show that perception of pseudo-V1CV2 can be obtained with formant transitions situated both inside and outside the vowel triangle, in which consonant C is one of three consonants /b, d, ɣ/, with synthesized acoustic transition and no burst. These results follows up the results previously published in Carré's deductive approach (2004) proposing a dynamic view of speech production, and on the prediction of vowel systems (Carré, 2009a): transition direction and length on frequency domain, and of transition rate (slope) in time domain can distinguish VV, CV stimuli (Carré, 2009b; Carré et al., 2007; Nguyen, 2009). We call them "acoustic vocalic gesture".

This performed hypothesis was based on formant frequency transition. As the description in section 1.1 (Chapter 1) about acoustics of speech production, formant frequency measurements are used in phonetics as part of the acoustic description of vowel like sound. However, formant frequencies are notoriously difficult to measure in a fully automatic way (Weenink, 2015). Many authors investigated the measurement methods of formant frequencies, such as linear prediction (Hunt, 1987; Schmid and Barnard, 1995), or analysis by synthesis with Fourier spectra (Welling and Ney, 1996), or peak picking on cepstrally smoothed spectra (Laprie and Berger, 1992). However, evaluting formant frequencies of speech is a technically challenge. Some obtained errors were analyzed, such as in (Atal and Schroeder, 1974; Monsen and Engebretson, 1983; Wood, 1989). They showed the difficulties in measuring formant frequencies related to (i) the ambiguous definition of the object to be measured; (ii) the spectral properties of the original signal that may be distorted by tape recorders, amplifiers and spectrograph; (iii) the spectrogram itself has very diffuse contours; and (iv) the measuring procedure.

In summary, it is difficult to measure formant frequencies value accurately during speech, especially consonantal production, and also with voice of high fundamental frequency such as female voice and child voice, see an example in poor sonogram of /bɤk/ pronounced by a Vietnamese female voice in Figure 2-1 in Chapter 2. From these points, it appears worthwhile to study and find out other parameters which can replace formant frequencies and act as "pseudo-formant" even during a consonant production. The objective of Chapter 4 is modeling acoustic and dynamic speech features with the similar approach on formant frequencies of the previous studies. And also due to the limitations in measuring of formant frequencies, we intend to do modeling acoustic and dynamic speech features based on "pseudo-formants" parameters. Therefore, two main parts will be carried out in this chapter, as follows:

- The first part presents on section 4.1, is to find out other parameters that can replace formants and act as "pseudo-formant" even during consonant production. This step is considered as a tool for the second part;
- In the second part, we study the way to be able to model acoustic and dynamic speech features from the "pseudo-formant" proposed in the first part.

4.1 The "pseudo-formant" parameters - Spectral Subband Centroid (SSC) features

Spectral Subband Centroid (SSC) features are considered as "pseudo-formant" parameters that have similar properties to formant frequencies (Paliwal, 1998). However, Paliwal only gave this comment without evidence. In our work, the SSC features will have small change with the original one of Paliwal and then give some comparisons between SSC features and formant frequencies in order to prove that SSC features are not formant frequencies, but are similar to formant frequencies.

4.1.1 Definition of SSCF features

Spectral Subband Centroid (SSC) was proposed by Paliwal (1998) and defined as follows:

- Step 1: Divide the frequency band (0 to $F_s/2$, where F_s is the sampling frequency in Hz) into a M fixed number of subbands. Each subband has lower and higher edges and a filter shape;
- Step 2: Compute the centroid for each subband using the power spectrum of the speech signal. Each centroid has its frequency and its magnitude.

Figure 4-1 illustrates the spectral subband centroid for one subband:

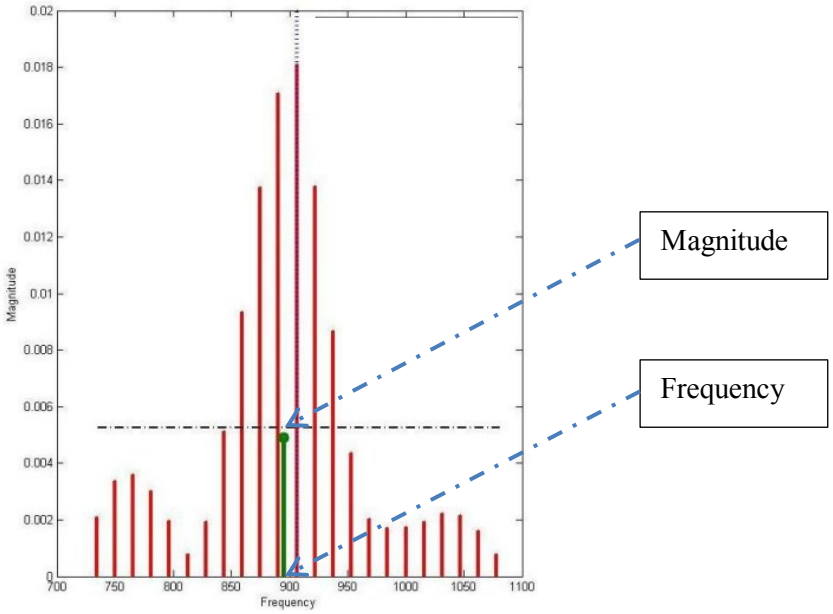


Figure 4-1: Subband signal, average energy (black dashed line), spectral subband centroid frequency (SSCF) and spectral subband centroid magnitude (SSCM).

The *m*th spectral subband centroid frequency $SSCF_m$ is computed by the following formula (4-1):

$$SSCF_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^{\gamma}(f) df}{\int_{l_m}^{h_m} w_m(f) P^{\gamma}(f) df} \quad \textbf{(4-1)}$$

where l_m and h_m are the lower and higher edges of *m*th subband, respectively; $w_m(f)$ is its shape; $P(f)$ is the power spectrum and γ is a constant controlling the dynamic range of the power spectrum.

SSCF is considered as formant-like feature (Paliwal, 1998), and these features can be extracted easily and reliably from the power spectrum of the speech signal.

The algorithm diagram computes SSCF parameters as the following figure 4-2. This algorithm is very simple.

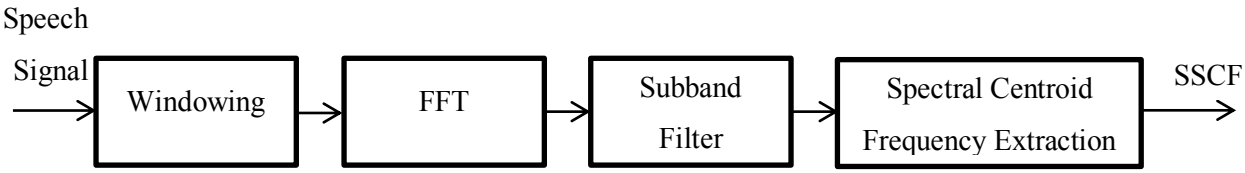


Figure 4-2: SSCF extraction algorithm.

4.1.2 Design of SSCF features

Kuldip K. Paliwal – who proposed the SSC features – chose to set the number of subband filters at M = 3 because in his view, these features were only used as supplementary for speech recognition (Paliwal, 1998).

According to our point of view, we would like to develop the dynamic approach which is not made of a static approach improved with dynamic parameters, but which consists in an intrinsically dynamic approach. Therefore we need more than three SSC parameters to help us more information on speech signal to model the dynamic parameters. Thus, in my thesis, we choose the number of subband filters at M = 6 (correspond to six pseudo-formants) and the filter bank is designed by dividing the frequency band uniformly on the mel scale with a triangle shape, as in Figure 4-3.

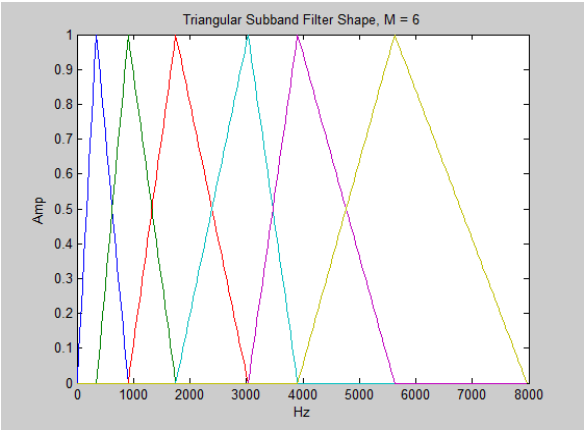


Figure 4-3: Subband filter shapes for computing SSCF with M = 6.

4.1.3 Comparison between SSCF features and formant frequencies

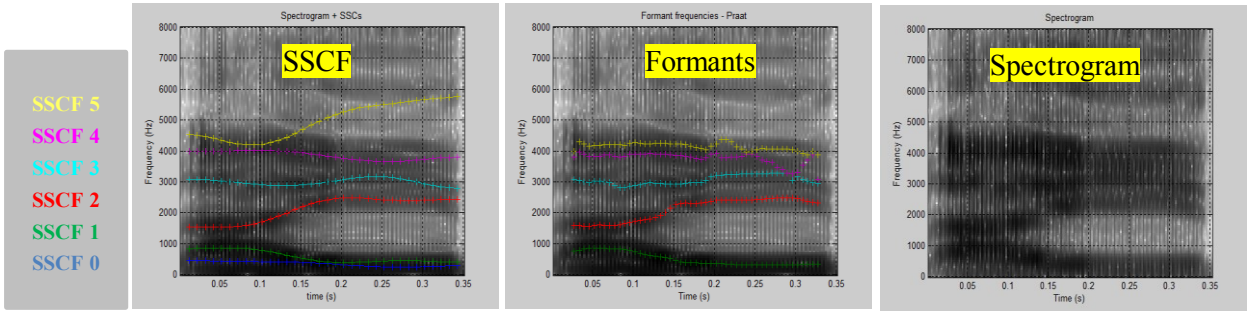
This comparison was performed on natural Vietnamese signal. SSCF was considered as formant-like features as in Paliwal's work (1998). We emphasize that SSCF is not formant frequencies here, thus all comparisons between two features are based on the shape, not on the value.

4.1.3.1 SSCF features have properties similar to formant frequencies

Figures 4-4 and 4-5 show one example: /ai/ stimuli of Vietnamese pronounced by one native male speaker and one native female speaker, respectively.

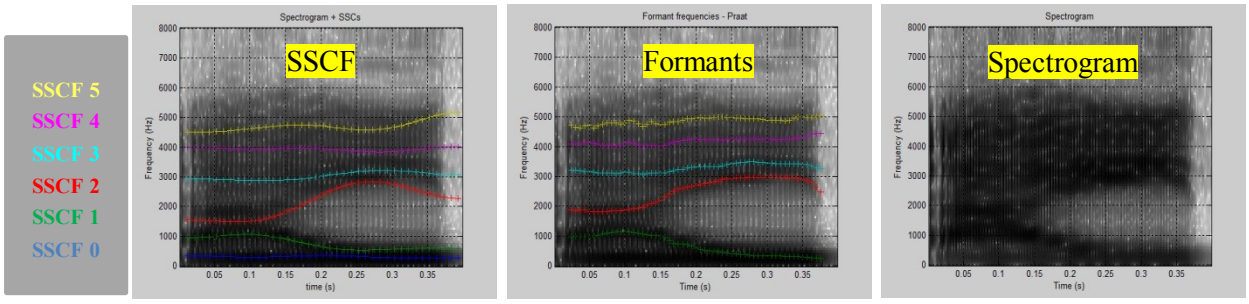
We give three figures of stimuli /ai/:

- the first one expresses SSCF parameters;
- the second one plots formant frequencies that are obtained by Praat toolkit;
- and the last one shows the spectrogram.



a) b) c)

Figure 4-4: /ai/ produced by one Vietnamese native male speaker: a) SSCF parameters (left); b) Formant frequencies (obtained from Praat toolkit); c) Spectrogram (right).



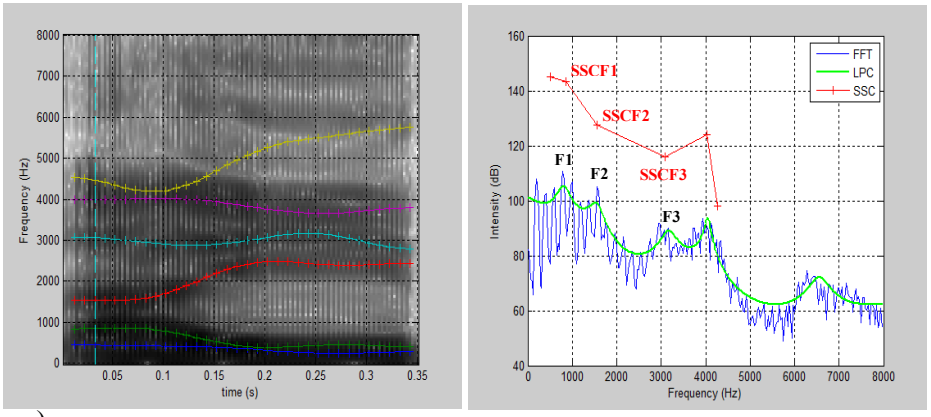
a) b) c)

Figure 4-5: /ai/ produced by one Vietnamese female speaker: a) SSCF parameters (left); b) Formant frequencies (obtained from Praat toolkit); c) Spectrogram (right).

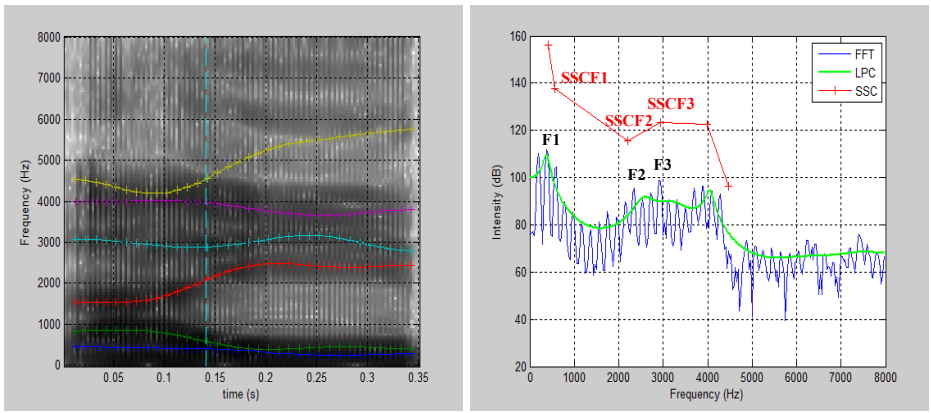
Observing Figure 4-4 for male voice, and compare the shapes among SSCF parameters (Figure 4-4-a), formant frequencies (Figure 4-4-b) and energies in spectrogram (Figure 4-4-c), we can see clearly that for /ai/ stimuli, SSCF parameters have similar shape with formant frequencies, especially SSCF1, SSCF2 follow very close F1, F2 formant frequencies. This comment still works for female voice in Figure 4-5.

In order to compare carefully SSCF parameters and formant frequencies, we illustrate three slices (that are plotted with cyan dashed lines) of /ai/ stimuli on spectra in Figure 4-6 produced by one male voice and in Figure 4-7 produced by a female voice that are corresponding to, respectively:

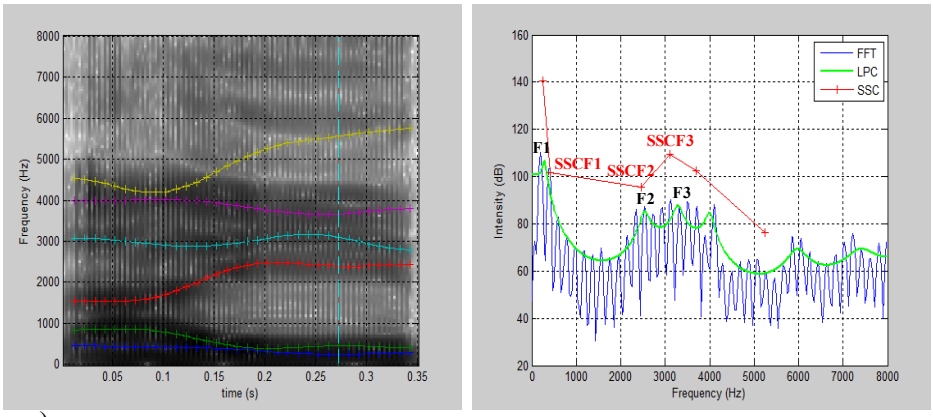
- one stable point of vowel /a/ (the top of the figure);
- one transition point of change section from vowel /a/ to vowel /i/ (the center figure) and;
- one stable point of vowel /i/ (the below figure).



a)



b)



c)

Figure 4-6: Spectra of /ai/ stimulus produced by one Vietnamese native male: a) *a stable point of vowel /a/ (top);b) a transition point of change section from vowel /a/ to vowel /i/; c) a stable point of vowel /i/ (bottom).*

In each spectra slice, in order to determine the formant frequency (i.e. F1, F2 and F3) point in each spectrum slice, we base on both spectrogram (blue color) and the peaks of LPC spectrum (green color). Because the peaks of LPC spectrum sometimes are not corresponding to formant frequencies, for example F3 in Figure 4-6-b for male voice or F2 in Figure 4-7-a for female voice. And we express each red plus point is corresponding to a SSCF parameter.

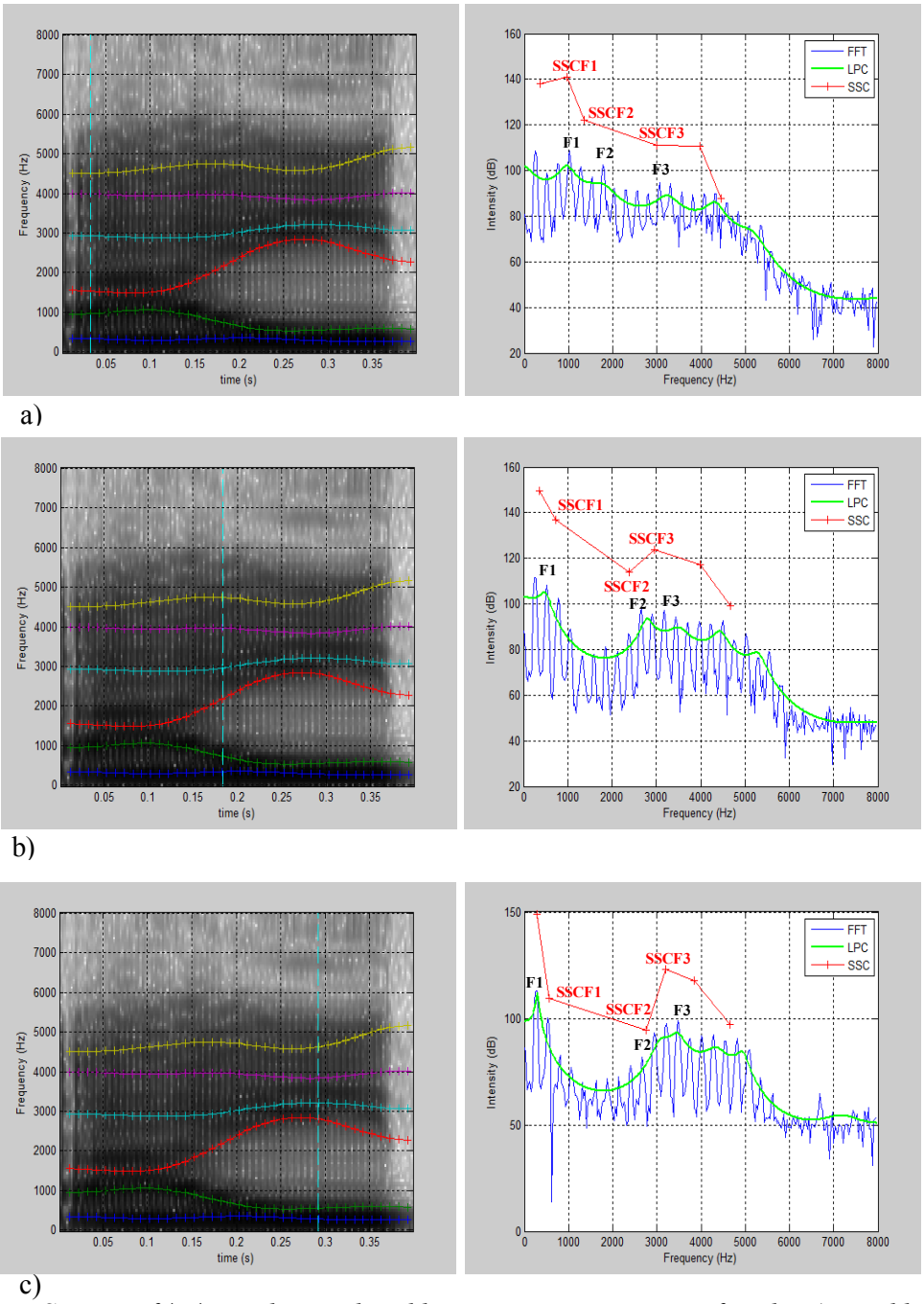


Figure 4-7: Spectra of /ai/ stimulus produced by one Vietnamese native female: a*) a stable point of vowel /a/ (top);b) a transition point from vowel /a/ to vowel /i/; c) a stable point of vowel /i/ (bottom).*

Obviously, SSCF and formant frequency points are very close position in time domain, especially SSCF1, SSCF2 are close with the position of F1, F2 formant frequencies.

This result suggests that SSCF parameters are very similar to formant frequencies for both male and female voice.

4.1.3.2 SSCF as continuous parameters on time domain, unlike formant frequencies

Obstruent productions have nearly no information on formant frequencies, thus formant frequencies are usually discontinuous tracking formant on time domain in real speech signal because it exits a lot of appearance of obstruent in syllables.

However on the contrary, SSCF parameters are only based on power spectrum of speech signal, thus it is easy and calculable to compute SSCF parameters even during consonant production. SSCF parameters are continuous lines on time domain. We see clearly this point when observing SSCF parameters and formant frequencies on the same two examples of fricatives: the consonant /s/ and /ʃ/ in the stimuli /asi/ in Figure 4-8 and /aʃa/ in Figure 4-9 that are pronounced by a Vietnamese native female speaker.

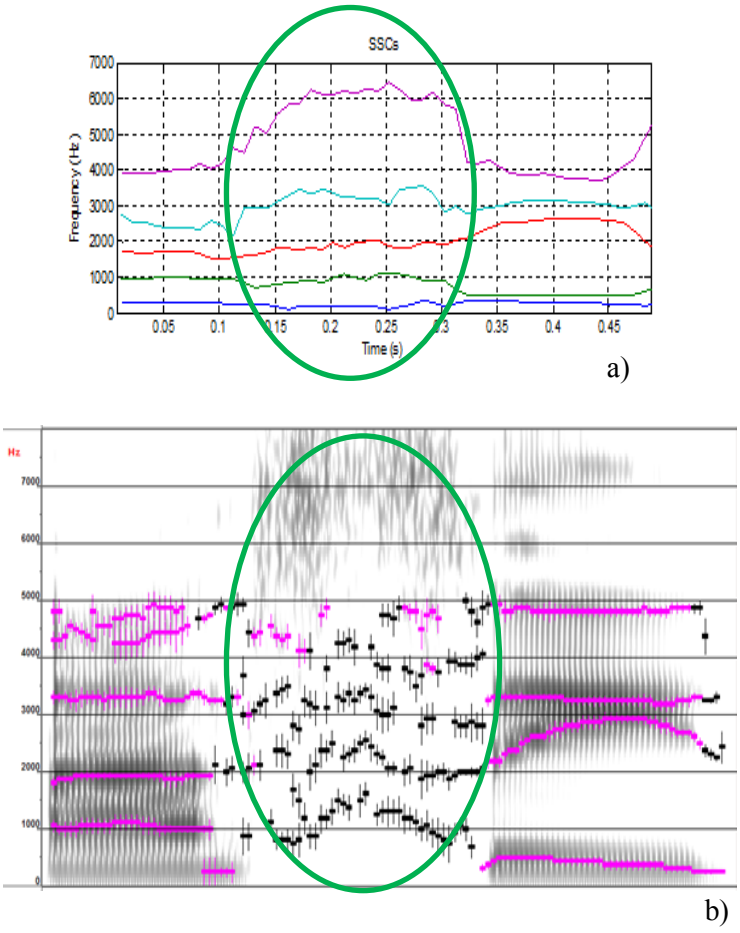


Figure 4-8: /asi/ of Vietnamese females: a) *SSCF parameters (top);*
b) formant frequencies (bottom: obtained from WinSnoori toolkit).

Figure 4-8 shows the productions by a native Vietnamese female of the unvoiced alveolar fricative /s/ that are marked in the green ellipse region in the /asi/ stimulus. We find that the lack of formant frequencies during the acoustic realization of the fricative /s/ is clearly visible explaining the

discontinuous tracking formant on the /asi/ stimulus, while SSCF parameters are continuous lines can be extracted during this consonant production.

In Figure 4-9, the invoiced alveopalatal fricative /ʃ/ produced by a native female speaker of Vietnamese and are marked by the green ellipse region in the /aʃa/ stimulus. As observed in the fricative /s/, SSCF parameters are continuous lines even during the production of consonant /ʃ/, while it is difficult to extract their formant frequencies.

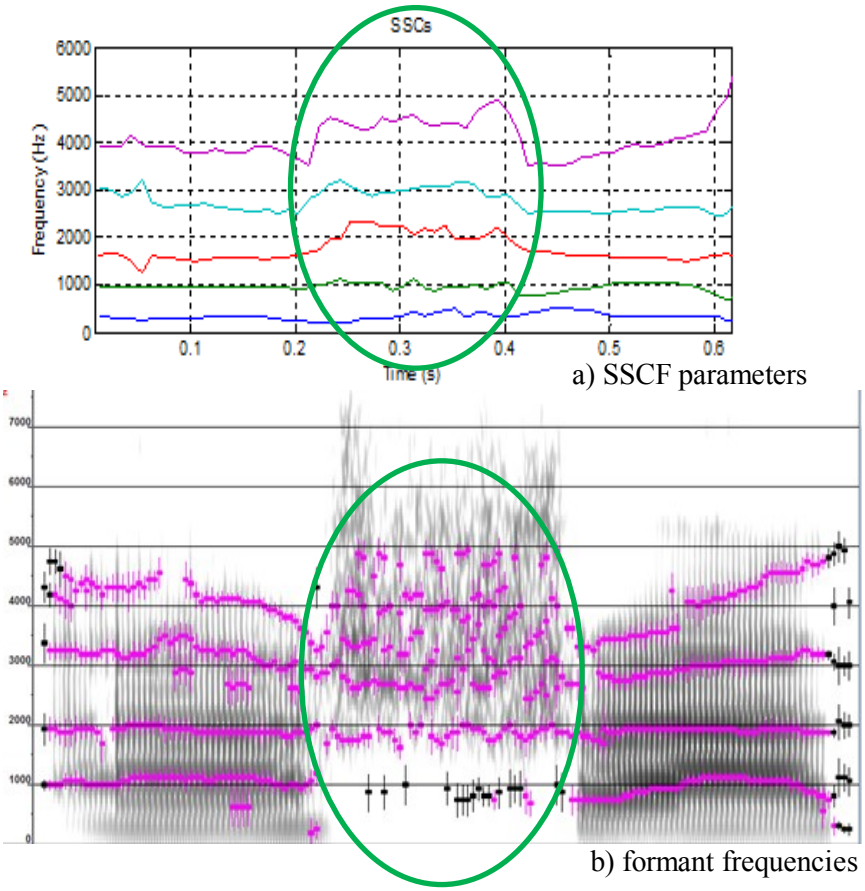


Figure 4-9: /aʃa/ of Vietnamese female: a) SSCF parameters (top); b) formant frequencies (bottom: obtained from WinSnoori toolkit).

Additionally, the comparison of the SSCF parameters between the two fricatives /s/ and /ʃ/ reveals that SSCF parameters of the alveolar consonant /s/ is distributed at higher frequencies than ones of the consonant /ʃ/. This point is completely suitable for natural distribution of the two consonants /s/ and /ʃ/ in the real speech signal, since in natural speech /s/ is always distributed at higher frequencies than the consonant /ʃ/ (Calliope Firm et al., 1989).

In conclusion of this part, contrary of formant frequencies, SSCF parameters are continuous parameters on time domain, even during consonant production.

4.1.3.3 Isolated vocalic SSCF parameters and vocalic formant frequencies

For real speech of each language, although everyone produce different formant values for the same vowels, vowels seem to lie within a specific region of formants F1/F2. In a recent study on Vietnamese, Castelli (2013) plotted nine Vietnamese vowels /a, ɛ, e, i, ɔ, o, u/ in F1-F2 plane. And the Vietnamese vocalic system is given in Figure 4-10.

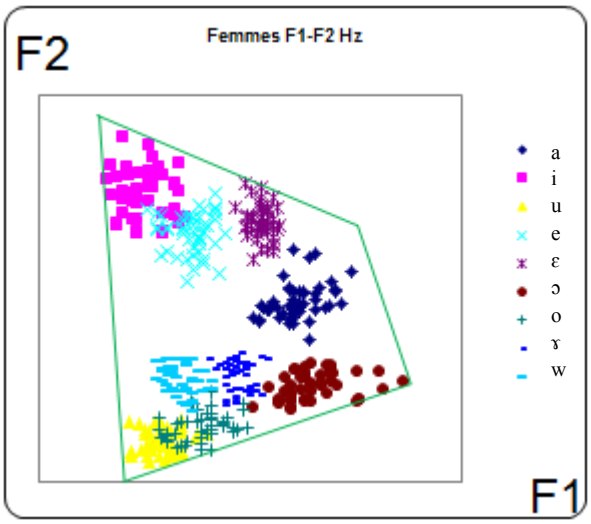


*Figure 4-10: Vocalic formant frequencies of 9 Vietnamese vowels /*a, *ɛ, e, i, ɔ, o, u, ɤ, w/.*

We study these nine Vietnamese vowels with SSCF by plotting them on the plane of SSCF1 and SSCF2. The results are shown in the below Figure 4-11:

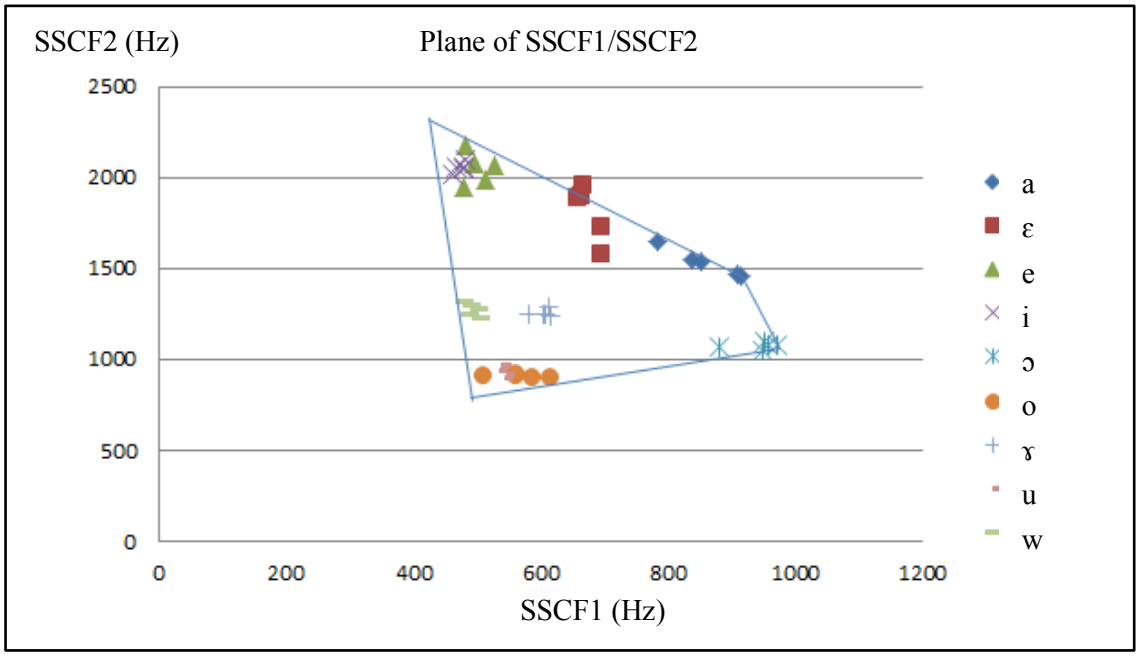


*Figure 4-11: Vocalic SSCF of 9 Vietnamese vowels /*a, *ɛ, e, i, ɔ, o, u, ɤ, w/.*

Figures 4-10 and 4-11 show that the shape of the vowel SSCF parameters is fairly similar to the one of vowel formant frequencies for Vietnamese.

Following up the obtained results in section 4.1.3.1, 4.1.3.2 and 4.1.3.3, we suggest that SSCF parameters of Vietnamese vowels are similar to its formant frequencies. However, opposite to formant frequencies, SSCF parameters are continuous on time domain, even during consonant production. Therefore, we suggest that SSCF parameters are "pseudo-formant" and can be used to replace formant frequencies in our following study on modeling acoustic and dynamic speech features.

4.2 Modeling acoustic dynamic speech features – SSCF Angles

Keeping up Alliot's result on acoustic Vietnamese vowel gesture on F1/F2 plane in section 1.5 (Chapter 1), we proposed that the acoustic vowel gesture can be defined with simple form: the straight line of formant frequencies and the ellipse shape of its formant speed. Therefore, if we can characterize the formant trajectories and its speed, we can characterize the acoustic gesture.

However, Alliot's algorithm for estimating formant frequencies is very complex, and it seems to be incomplete for female voices. Meanwhile, with the description of SSCF parameters and, together with the comparison results between SSCF parameters and formant frequencies presented in section 4.1, SSCF parameters are considered to be similar to formant frequencies. Furthermore, the algorithm of SSCF is simply and easy to deploy.

Consequently, a question is raised here: can the SSCF trajectories and their speeds make a similar acoustic vowel gesture to the one of formant frequencies plotted in Figure 1-19 in section 1.5.3 (Chapter 1)? The answer will be given in the following section.

4.2.1 Acoustic Vietnamese vowel gesture on SSCF parameter plane

4.2.1.1 Methodology

4.2.1.1.1 Stimuli

We studied on six Vietnamese vowel-to-vowel stimuli /ai, ia, au, ua, ui, iu/. Each item was recoded ten times (5 times at normal rate and 5 times at fast rate) by two native Vietnamese speakers (1 male and 1 female) from the north of Vietnam with age from 25 to 32 years old.

4.2.1.1.2 Implementation

For our study, we calculated SSCF parameters (following the algorithm presented in section 4.1.1) of six vowel-to-vowel stimuli in section 4.2.1.1.1 and performed them in the SSCF1/SSCF2 plane and the plane of the speeds of SSCF1 and SSCF2.

4.2.1.2 Results

Figure 4-13 and 4-14 show results on SSCF1/SSCF2 plane and on their speed plane for one native Vietnamese male speaker and for one native Vietnamese female speaker, respectively.

In order to easily compare with Alliot's result on formant vowel gesture in section 1.5 (Chapter 1), we get a copy of Figure 1-19 into following Figure 4-12.

Analyzing the results plotted in these two Figures 4-13 and 4-14, allows to see that firstly, on the SSCF1/SSCF2 plane, all SSCF trajectories are fairly straight lines. And three pairs of transition can be obviously separated (on the left of Figure 4-13 for Vietnamese male voice, and on the left of Figure 4-14 for Vietnamese female voice): first pair is a group of /ai, ia/ that is expressed by blue solid or plus lines, respectively; second pair is a group of /au, ua/ that is expressed by green solid or plus lines, respectively; and third pair is a group of /ui, iu/ that is plotted by red solid or plus lines, respectively. This result is similar to one obtained by Alliot (2009) for formant trajectories in F1/F2 plane in the left of Figure 1-19 (or Figure 4-12).

On the SSCF1/SSCF2 speed plane, the SSCF speed trajectory of each transition has also more or less ellipse shape. The six transition types correspond to /ai, ia, au, ua, ui, iu/ can be well visually discriminated (on the right of Figure 4-13 for Vietnamese male voice, and on the right of Figure 4-14 for Vietnamese female voice). It is found that the speed direction of each transition is fairly opposite to one of its inverse transition (example for /ai/ and /ia/; or /au/ and /ua/; or /ui/ and /iu/). This result is similar to one obtained by Alliot (2009) for formant speed trajectories on F1 speed/F2 speed plane.

SSCF results for both Vietnamese male and female show that the one of male voice is better than the one for female voice. This can be explained by the effect of poor harmonic in high voices. However, we can still separate six transitions of /ai, ia, au, ua, ui, iu/ based on the direction and speed of SSCF for both Vietnamese male and Vietnamese female.

Comparing visually the acoustic vowel gestures obtained from formant frequencies and those from SSCF parameters (Figure 4-12, to 4-14) points out that the shapes of SSCF transition trajectories in SSCF1/SSCF2 plane and in the speed plane are fairly similar to those of formant transition trajectory in F1/F2 plane and in the corresponding speed plane, respectively.

The next figures show the transition trajectories of the six targets /ai, ia, au, ua, ui, iu/ (produced by the same speaker at figures 4-13 and 4-14, respectively) on 3-D plane of SSCF1/SSCF2/SSCF3 and the six corresponding SSCF speed trajectories on 3-D plane SSCF1 speed/SSCF2 speed/SSCF3 speed.

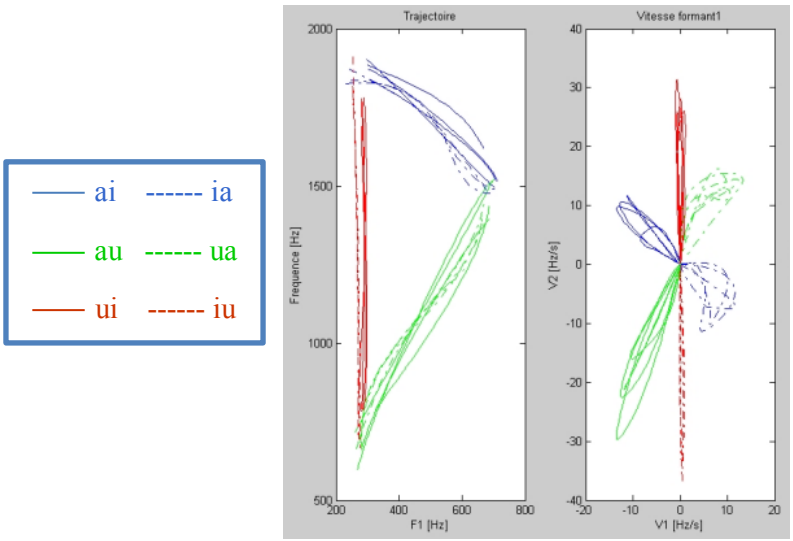


Figure 4-12: Vowel-to-Vowel transitions in the F1/F2 plane (left) and transition speeds plane produced by a native male speaker of Vietnamese (Alliot, 2009) (copy from Figure 1-19).

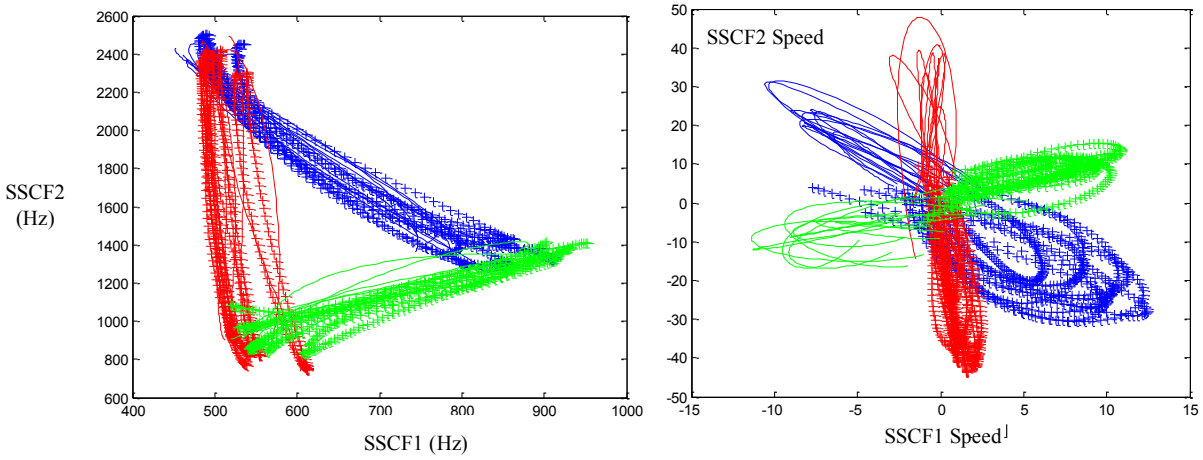


Figure 4-13: Vowel-to-Vowel transitions on SSCF1/SSCF2 plane and transition speeds produced by a native male speaker of Vietnamese.

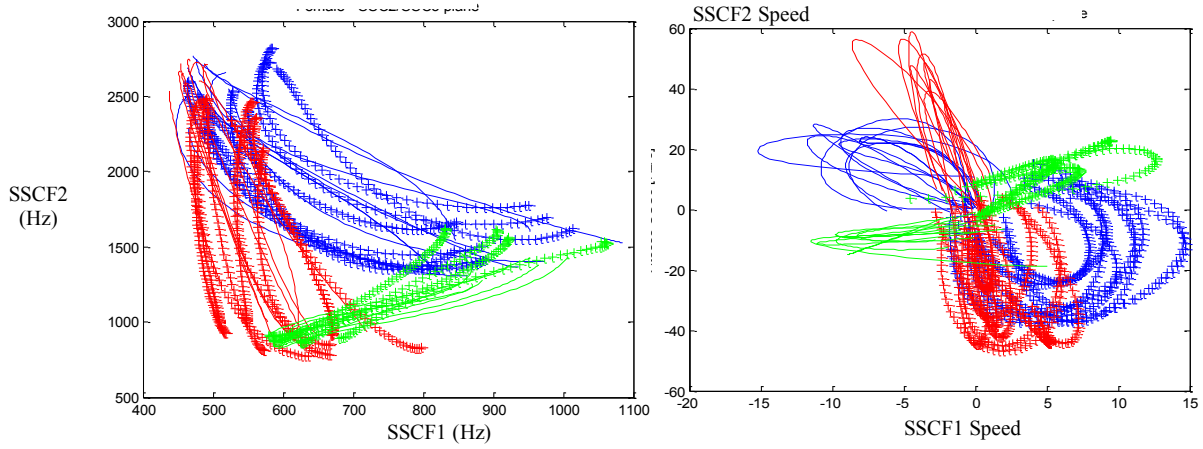


Figure 4-14: Vowel-to-Vowel transitions on SSCF1/SSCF2 plane and transition speeds produced by a native female speaker of Vietnamese.

The results make more clearly the separation between the six transitions based on the 3-D direction and the corresponding 3-D speed plane of SSCF.

In the same way as results in 2-D plane, and giving the information on acoustic gestures, all SSCF trajectories are fairly straight lines, and the SSCF speed trajectory of each transition have more or less ellipse shape.

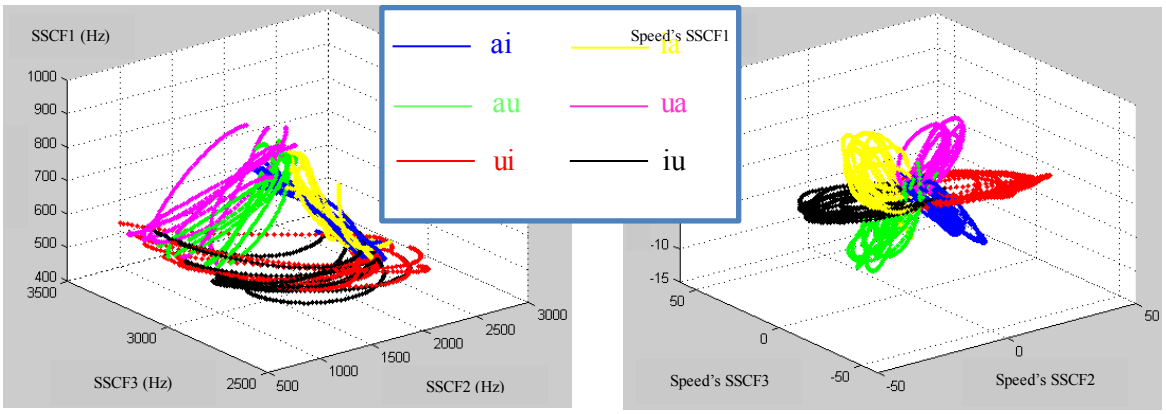


Figure 4-15: Vowel-to-Vowel transition in 3-D plane SSCF1/ SSCF2/SSCF3 (left) and the corresponding transition speed 3-D space (right) produced by a native Vietnamese male speaker.

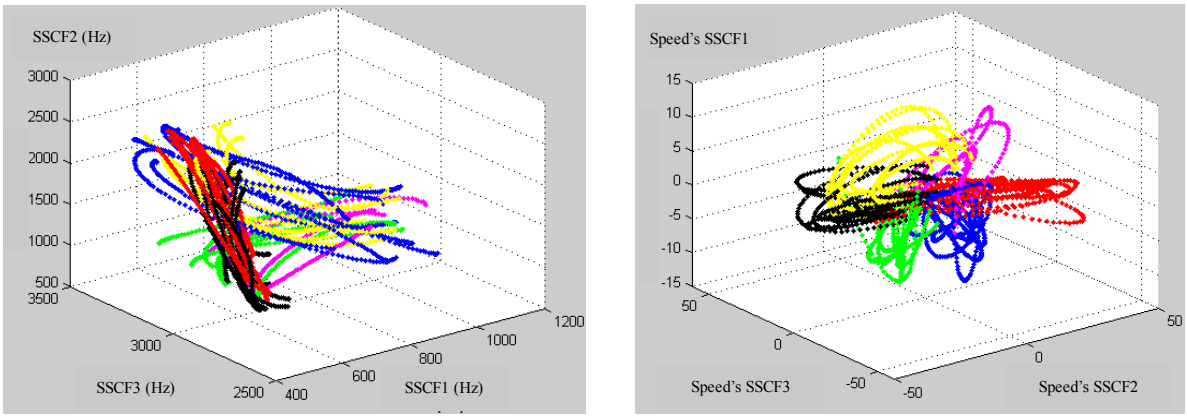


Figure 4-16: Vowel-to-Vowel transition in 3-D plane SSCF1/ SSCF2/SSCF3 (left) and the corresponding transition speed 3-D space (right) produced by a native Vietnamese female speaker.

In summary, it seems that SSCF parameters can replace formant frequencies and act as "pseudo-formants". Consequently, the acoustic gesture can be defined with a simple form: the straight lines of SSCF trajectories and the ellipse shapes of their SSCF speeds. Therefore, it can be suggested that if we can characterize the SSCF trajectories and their speeds, then we can characterize acoustic gesture.

4.2.2 Modeling acoustic and dynamic speech features from SSCF parameters – SSCF Angles

From the previous results presented in section 4.2.1, we proposed that acoustic gestures were defined by SSCF trajectories and SSCF speed with the following characteristics:

- all SSCF trajectories are fairly straight lines;
- the SSCF speed trajectory of each transition has more or less ellipse shape.

We proposed that the dynamic speech features here are SSCF transition directions: in other words, they can be modeled by SSCF Angles in SSCF plane (assuming all SSCF trajectories are more or less straight lines). SSCF Angles can be defined as angles between the starting and the ending points of speech transition trajectories in corresponding SSCF plane.

An example on the SSCF1/ SSCF2 plane in Figure 4-18, the SSCF Angles12 are angles between the starting and ending points of V1V2 transition trajectories in the SSCF1/SSCF2 plane, and are signed by parameter α.

On each $SSCF_i$ / $SSCF_{i+1}$ plane , the formula of the SSCF Angle(i)(i+1) is defined as:

$$\text{angle(i)(i+1)} = \text{atan}\left(\frac{\Delta SSCF_{i+1}}{\Delta SSCF_i}\right) \text{ in degree } (^o) \qquad \textbf{(4-2)}$$

Where:

$\Delta SSCF_{i+1}$ is the difference between $SSCF_{i+1}$ at the end and at the beginning of the transition.

$\Delta SSCF_i$ is the difference between $SSCF_i$ at the end and at the beginning of the transition.

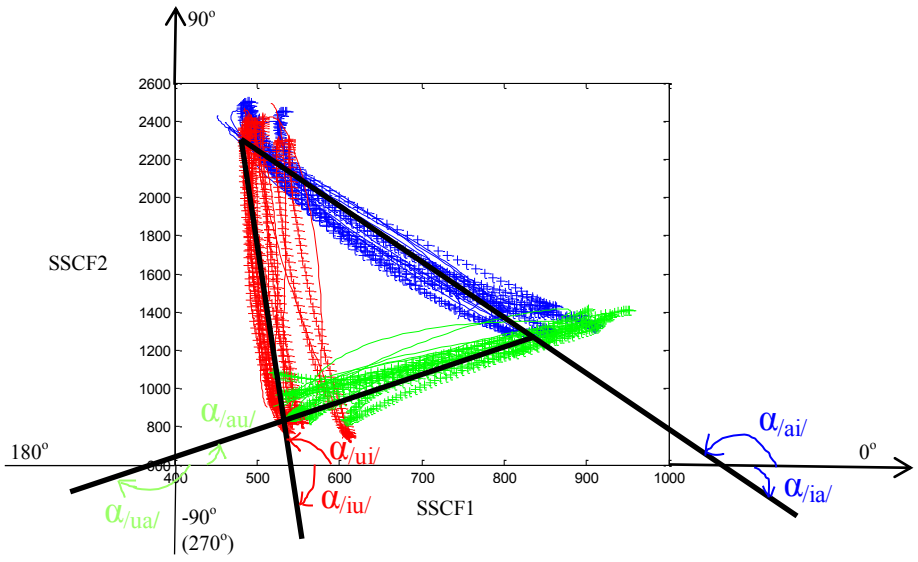
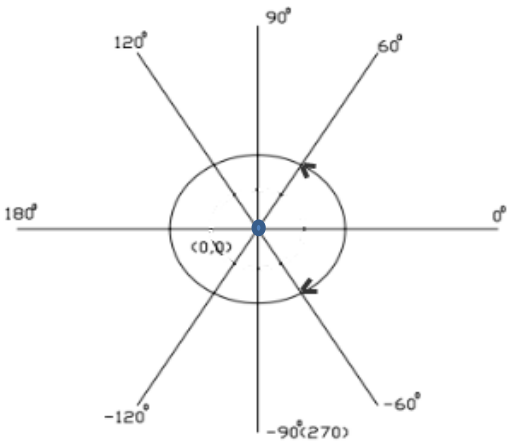


Figure 4-18: SSCF Angles12 in SSCF1/SSCF2 plane.

Figure 4-17: Angle values in the polar coordinate system.



Obviously, to be able to compute SSCF Angles, we needed to determine a part that is more or less transition part with its beginning and ending point. This point will be solved in section 4.3.

In the polar coordinate system, the angles vary from -180^o to +180^o (or 0^o to 360^o), as presented in Figure 4-17.

4.3 Calculation of the acoustic and dynamic speech features using SSCF Angles

As in section 4.2.2, we proposed SSCF Angles as the dynamic speech features that are computed from SSCF parameters. Therefore, in order to study step by step, in our thesis study, we focus on SSCF Angles analysis on Vowel-to-Vowel (V1V2) transitions, and SSCF Angles on Vowel-Consonant-Vowel (VCV) transition will be considered as part of the perspectives for future work. Consequently, the SSCF Angles on V1V2 transition are computed by the following method.

Each V1V2 item will be defined by six SSCF parameters: from SSCF0 to SSCF5. Each pair of these SSCFs will have one corresponding angle.

In order to compute angle parameter, we have to determine one middle part in each speech item so that this part is fairly close to the transition part in each item. That means we have to determine one beginning point and one ending point for this part.

The determination process of SSCF Angles for each V1V2 item is performed in three steps, as follows:

Step 1: Determines the middle point of the equivalent-transition part.

The estimation of the middle point is based on the speed of SSCF2. Depending on each type of transition item, the maximum or minimum of SSCF2's speed will be chosen as a criterion to determine the middle point of this equivalent-transition part. The time point corresponding to the middle point of this part is t_m.

- if SSCF2 parameters increase from V1 to V2, then the maximum of SSCF2's speed will be chosen as the determination criterion to determine the middle point of this equivalent-transition part, for example in case of /ai/, /aɛ/, /ae/, /ui/, /ua/, etc;
- on the contrary, if SSCF2 parameters decrease from V1 to V2, then the minimum of SSCF2's speed will be chosen as the determination criterion to determine the middle point of this equivalent-transition part, for example in case of /ia/, /ɛa/, /ea/, /iu/, /au/, etc.

Figures 4-19 and 4-20 show two examples of SSCF parameters and SSCF2's speed in /ai/ and /ia/ transition produced by a native Vietnamese male speaker. The vertical blue dashed lines show the maximum of SSCF2's speed in /ai/ and the minimum of SSCF2's speed in /ia/, respectively.

Step 2: Determines the beginning and the ending time points of the equivalent-transition part.

Keeping up the middle point of the equivalent-transition part determined in step 1, the corresponding beginning and the ending points of this part are defined as the following formulas:

$$\text{the beginning time point} = (1 - \text{threshold}) * t_m \quad \textbf{(4-3)}$$

$$\text{the ending time point} = (1 + \text{threshold}) * t_m \quad \textbf{(4-4)}$$

We choose threshold = 0.4. Two examples in Figures 4-19 and 4-20 continue to illustrate this step. According to the blue middle point and the two formulas (4-3), (4-4), the beginning points are expressed by the plus black points, and the ending points are expressed by the circle black points in both cases of /ai/ and /ia/ transition.

Step 3: After determining the beginning and the ending point of the equivalent-transition segment, we perform to calculate three angles (angle12, angle23 and angle34) by the formula (4-2) presented in the section 4.2.

After calculating the SSCF Angles, we continue to perform SSCF Angles analysis on some Vowel-to-Vowel (V1V2) transitions of both Vietnamese and French in order to assess the role of these parameters in speech discrimination. This work will be given in section 4.4.

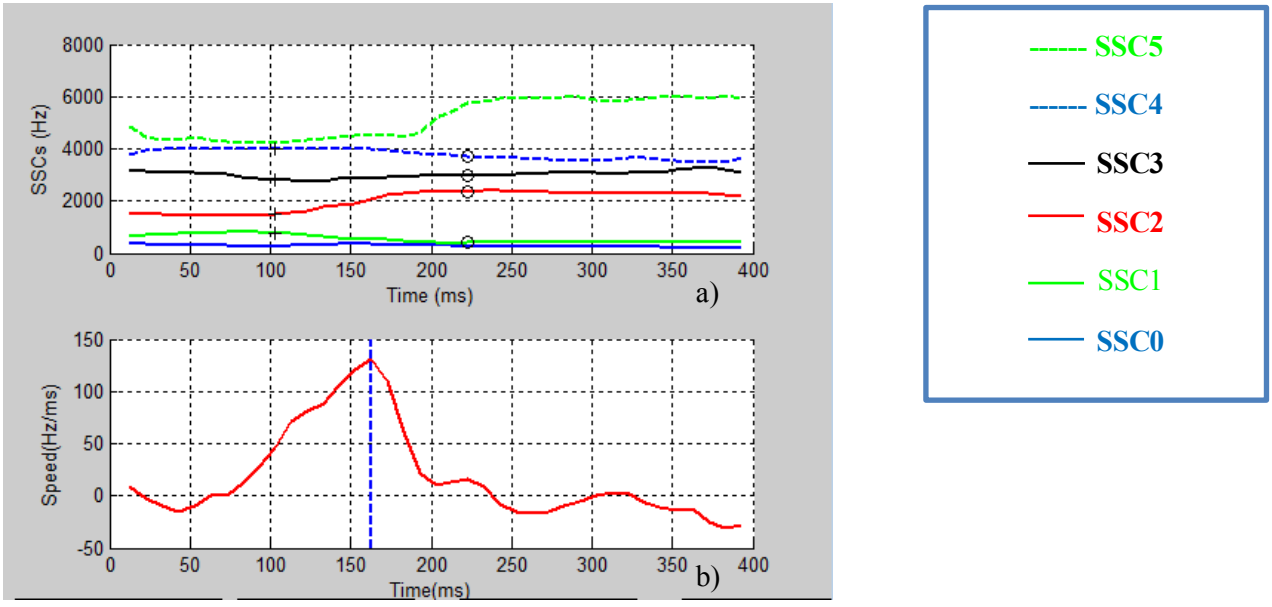


Figure 4-19: SSCF parameters (top) and SSCF2's speed (bottom) for /ai/ transition produced by a native Vietnamese male.

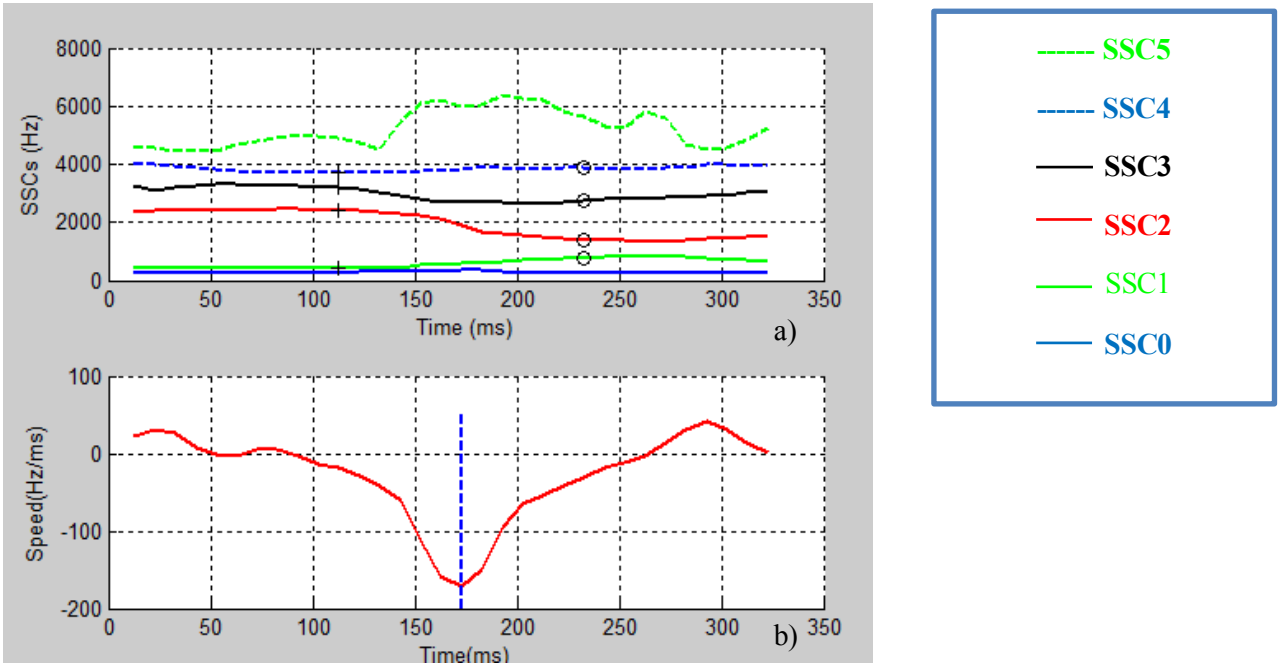


Figure 4-20: SSCF parameters (top) and SSCF2's speed (bottom) for /ia/ transition produced by a native Vietnamese male.

4.4 SSCF Angles analysis on Vietnamese Vowel – to – Vowel transitions

4.4.1 Methodology

4.4.1.1 Vietnamese stimuli

In order to study the Vietnamese Vowel-to-Vowel transitions, a small Vietnamese V1V2 transition corpus was built by eight Vietnamese native speakers (4 males and 4 females) who were born and live in the north of Vietnam, with ages from 25 to 32 years old. The below fourteen transitions /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ were studied. Each V1V2 sequence was inserted in the carrier sentence: "Nói V1V2 ba lần" [noj^{5} V1V2 ba^{1} lɤn^{2}] ("Speak V1V2 three times").

Each V1V2 sequence was produced ten times at two different speech rates (5 times at normal rate and 5 times at fast rate). Table 4-1 shows in detail the total studied sentence number in Vietnamese. There were 1.120 recording sentences for eight Vietnamese speakers.

The speaker was instructed to read the V1V2 sequences so that the transition from V1 vowel to V2 vowel is continuous in time domain. The recording process was controlled by PC software that randomly presented the succession of the items to be recorded. In case of bad pronunciation or hesitation, the speaker had to pronounce the item again.

This recording took place in quiet studio of Mica Institute, Hanoi, Vietnam. The corpus was recorded by microphone with a sampling frequency of 16.000Hz and 16bits per sample and save in .wav format.

The speaker was invited to read the corpus at normal speed. A small break was made during reading the different sentences. After completing the corpus at normal speed, the speaker was invited to read the corpus at fast speed. This process was going on repetitively for ten times of the recording.

These recorded data were segmented and extracted all sequences of V1V2 for analyzing.

Table 4-1: Total number of the studied sentences in Vietnamese.

	Number of items/speaker	Recording times/item/sentence		Total studied sentence number
		At normal rate	At fast rate	
1 speaker	14	5	5	140 sentences/speaker
8 speakers	14	5	5	1120 sentences/8 speakers

4.4.1.2 Analysis method

In order to point out the role of SSCF Angles in V1V2 sequence discrimination, for each V1V2 sequence, three SSCF Angles were computed, as follows:

- SSCF Angle12 is angle of V1V2 trajectory in the SSCF1/SSCF2 plane;
- SSCF Angle23 is angle of V1V2 trajectory in the SSCF2/SSCF3 plane;
- SSCF Angle34 is angle of V1V2 trajectory in the SSCF3/SSCF4 plane.

And then, these three angles were compared in two following aspects:

- first case is the SSCF Angles comparisons among different items for each speaker;
- second case is the SSCF Angles comparisons with the same item among males and females.

In each case, the SSCF Angles are also compared at between normal and fast speech rate for same item and same speaker.

All the analysis results will be presented in section 4.4.2.

4.4.2 Results

In the following results, each SSCF Angle for each participant will have two results:

- one result is for normal rate;

- one result is for fast rate.

Each SSCF Angle result at each speech rate (normal/fast) is presented by the angle average and its standard deviations. Each item was produced five times at normal rate and five times at fast rate.

4.4.2.1 Case 1: SSCF Angles comparisons among different transitions for each speaker

For this comparison, we will show the typical results with one Vietnamese male M1 and one Vietnamese female F1. The other results of other Vietnamese males and females will be presented in Appendix 1.

4.4.2.1.1 SSCF Angle12

Figures 4-21 and 4-22 present the average and standard deviation of SSCF Angle12 for the fourteen transitions /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ (obtained from the database of section 4.4.1.1) produced by one Vietnamese native male M1 and one Vietnamese native female F1, respectively.

Figure 4-21 shows the result of SSCF Angles of one Vietnamese male participant M1. Firstly, for each V1V2 transition sequence, the averages of SSCF Angle12 are more or less the same values and its standard deviation is small at both normal and fast rate. Secondly, the sums of the absolute of the SSCF Angle12 of each V1V2 item and the one of V2V1 sequence are fairly equal to 180^o. These mean that there is a symmetric property between V1V2 and V2V1 on acoustic SSCF Angle12 domain, on other words, in SSCF1/SSCF2 plane, the trajectories of V1V2 is fairly parallel with the ones of V2V1. Thirdly, six groups of V1V2 sequences are separated according to their different SSCF Angle12 values: group 1 includes the transitions of /ai, aɛ, ae/; group 2 is the /ui/ transitions; group 3 contains /ua, ɔa, oa/ transitions; group 4 involves /ia, ɛa, ea/ transitions; group 5 is the /iu/ transitions; and group 6 consists of the transition of /au, aɔ, ao/. The SSCF Angle12 values of the three first groups of V1V2 are positive, on the contrary, the ones of the three last groups of V2V1 are negative. This result is completely suitable for the above symmetric property of VV transitions on SSCF Angle12. These similar results were obtained for one Vietnamese female F1 in Figure 4-22. These characteristics are also observed for the other participants in Appendix 1.

In summary, we observe similar results between normal and fast production and the six groups of VV sequences can be discriminated according to their SSCF Angle12.

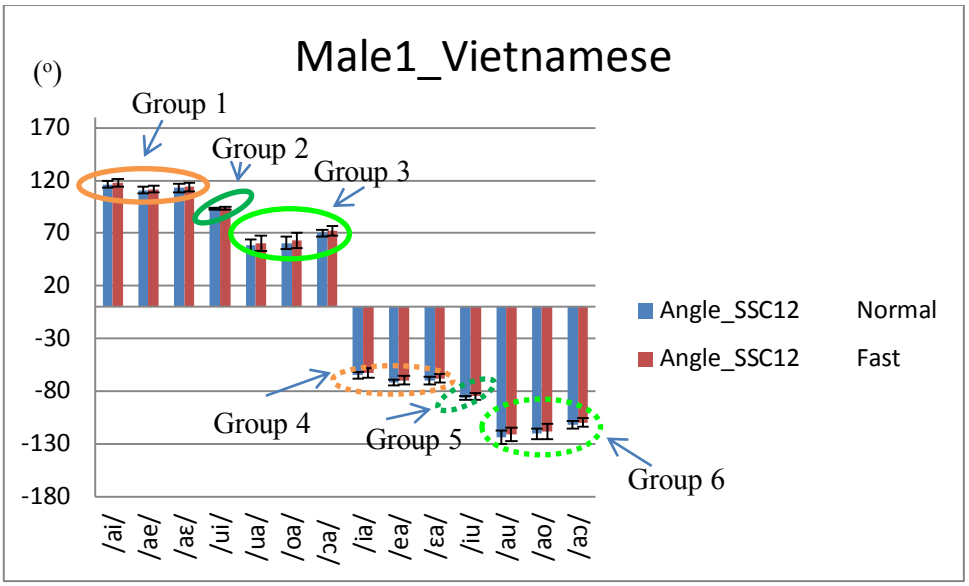


Figure 4-21: The average value and standard deviation of SSCF Angle12 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produced by one Vietnamese male (M1) at both normal and fast rate.

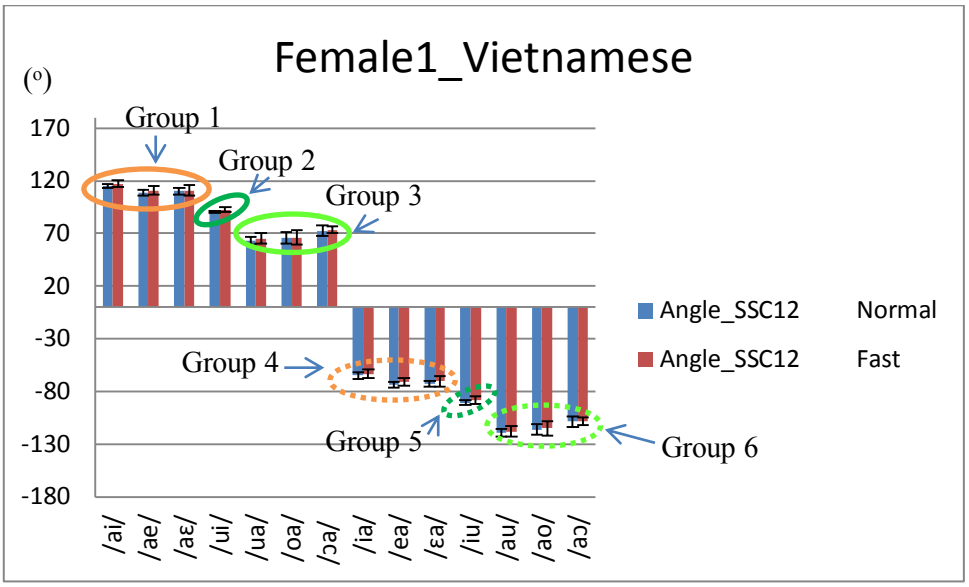


Figure 4-22: The average value and standard deviation of SSCF Angle12 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produced by one Vietnamese female (F1) at both normal and fast rate.

4.4.2.1.2 SSCF Angle23

The average and standard deviation values of SSCF Angle23 of the fourteen transitions /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produced by the same one male M1 and one female F1 are presented in Figure 4-23 and 4-24, respectively.

We point out similar comments for both Vietnamese male M1 and Vietnamese female F1. Firstly, for each item (/ai/, /aɛ/, /ae/, /ua/, /ɔa/, /oa/, /ui/, /ia/, /ɛa/, /ea/, /au/, /aɔ/, /ao/ or /iu/), the average values of SSCF Angle23 are more or less the same at both normal and fast rate with its small standard

deviation for each speaker. This indicates the invariant property of SSCF Angle23 of the same VV sequences produced by the same speaker at the different rate.

Secondly, the symmetric property between V1V2 and V2V1 on acoustic SSCF Angle23 domain are showed by the sums of the absolute value of the SSCF Angle23 of each V1V2 item and the one of each V2V1 sequence are fairly equal to 180°.

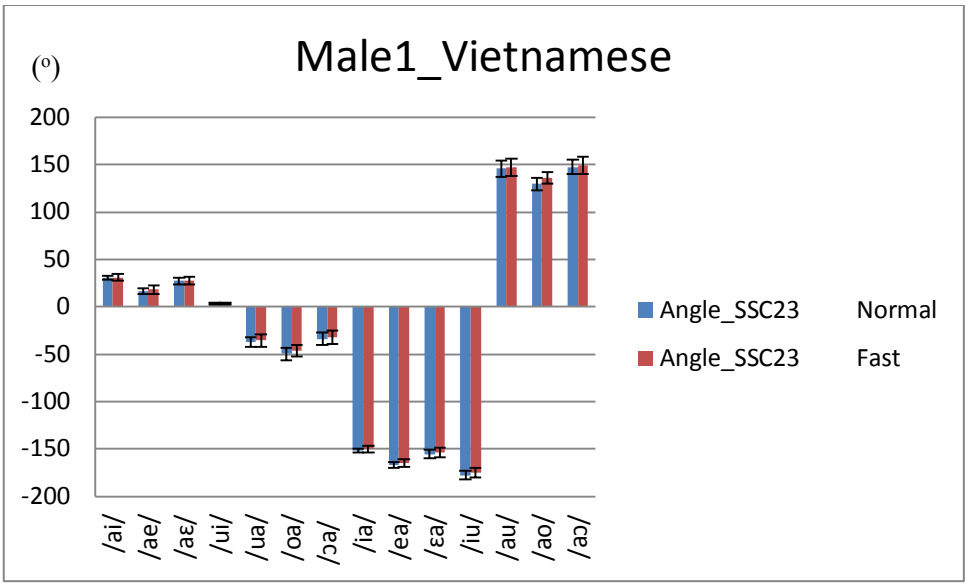


Figure 4-23: The average value and standard deviation of SSCF Angle23 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produced by one Vietnamese female (M1) at both normal and fast rate.

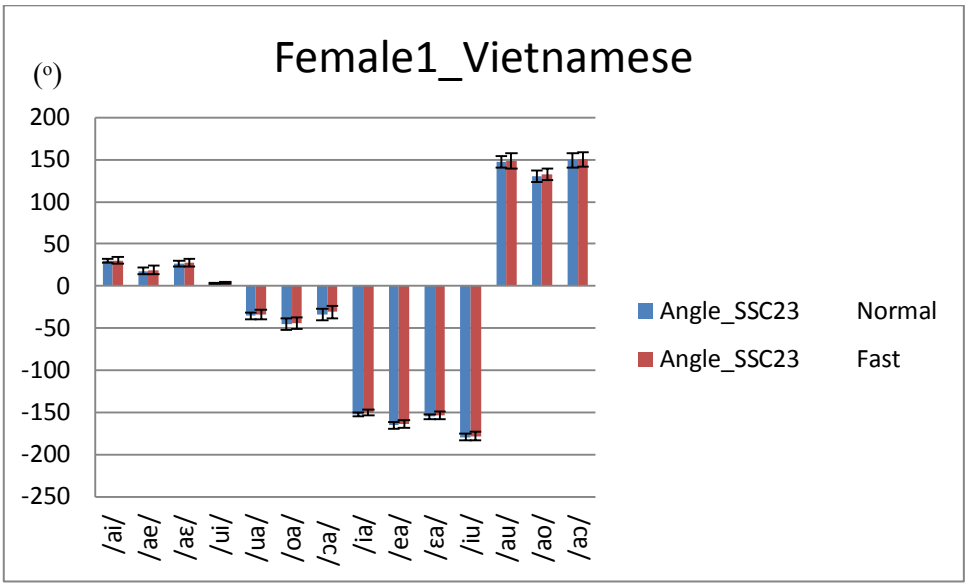


Figure 4-24: The average value and standard deviation of SSCF Angle23 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produced by one Vietnamese female (F1) at both normal and fast rate.

Thirdly, the SSCF Angle23 values divided all fourteen transitions into six separated groups: group 1 of /ai, ae, aɛ/; group 2 of /ui/; group 3 of /ua, oa, ɔa/; group 4 of /ia, ea, ɛa/; group 5 of /iu/ and group 6 of /au, ao, aɔ/.

All these comments are also obtained for the other speakers in Appendix 1. We observe the similar results between normal and fast production and the six separated groups of VV sequences can be discriminated according to their SSCF Angle23. This result is similar to the one of SSCF Angle12.

4.4.2.1.3 SSCF Angle34

Figure 4-25 and 4-26 present the average and standard deviation of SSCF Angle34 for the fourteen transitions /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ (obtained from the database of section 4.4.1.1) produced by the same male speaker M1 and female speaker F1, respectively.

For both Vietnamese male M1 and Vietnamese female F1, we observe the similar results. Firstly, for each item (/ai/, /aɛ/, /ae/, /ua/, /ɔa/, /oa/, /ui/, /ia/, /ɛa/, /ea/, /au/, /aɔ/, /ao/ or /iu/) of each speaker, the average of SSCF Angle34 are more or less the same value at both normal and fast rate with its small standard deviation. Secondly, like SSCF Angle12 an SSCF Angle23, the symmetric property between V1V2 and V2V1 are also presented on acoustic SSCF Angle34 domain by the sums of the absolute value of the SSCF Angle34 of each V1V2 item and the one of each V2V1 sequence are fairly equal to 180°. However, the SSCF Angle34 values are completely different for these fourteen transitions.

These same results on SSCF Angle34 are also observed for other speakers in Appendix 1. Obviously, the fourteen Vietnamese VV sequences were completely separated according to their SSCF Angle34 at both normal and fast rate.

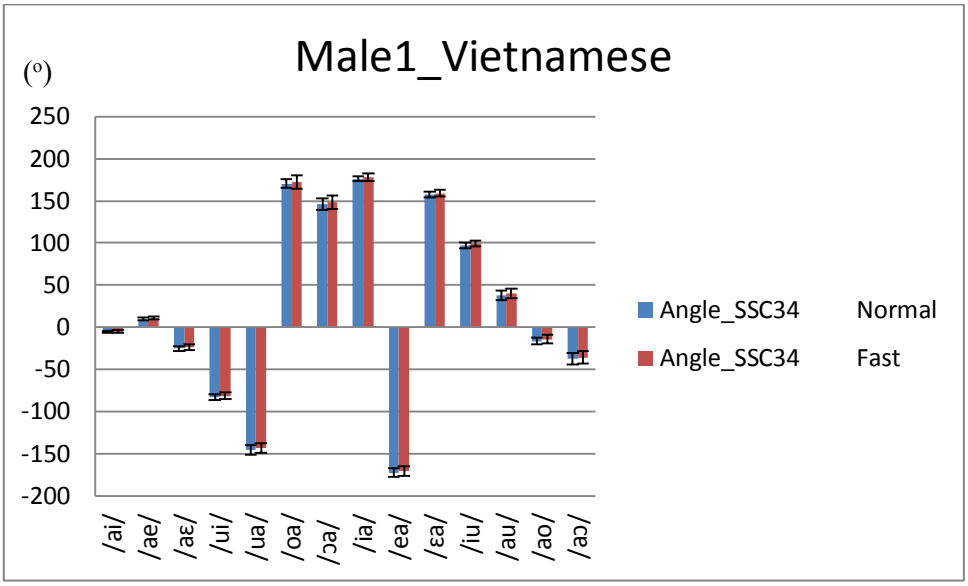


Figure 4-25: The average value and standard deviation of SSCF Angle34 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produced by one Vietnamese female (M1) at both normal and fast rate.

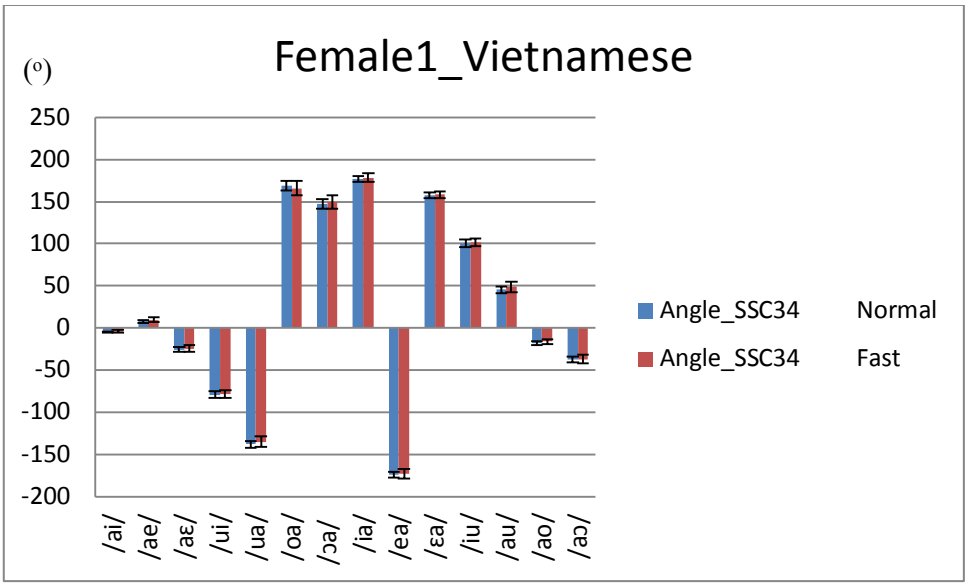


Figure 4-26: The average value and standard deviation of SSCF Angle34 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produced by one Vietnamese female (F1) at both normal and fast rate.

In summary, in this case of SSCF Angles comparisons among different items for each speaker, keeping up the results of SSCF Angle12, SSCF Angle23 and SSCF Angle34, there is similar result between normal and fast production for each SSCF Angle in each transition sequence produced by each speaker. However, while there are only six separated groups if based on SSCF Angle12 or SSCF Angle23: group 1 includes the transitions of /ai, aɛ, ae/; group 2 is the /ui/ transitions; group 3 contains /ua, ɔa, oa/ transitions; group 4 involves /ia, ɛa, ea/ transitions; group 5 is the /iu/ transitions; and group 6 consists of the transition of /au, aɔ, ao/, these fourteen transitions /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ were completely separated based on their SSCF Angle34.

4.4.2.2 Case 2: SSCF Angles comparisons with same items among males and females

Our first Vietnamese results on case 1 were presented the differences of fourteen Vietnamese VV productions produced by each Vietnamese speaker. In this case, we consider each VV item produced by eight Vietnamese native speakers.

4.4.2.2.1 /ai/ sequence

a, SSCF Angle12 of /ai/

Figure 4-27 presents the average results and its standard deviation of SSCF Angle12 of the /ai/ transition produced by eight Vietnamese speakers (four males and four females). Some observations are given. Firstly, for each speaker, the average values of SSCF Angle12 of /ai/ transition are positive and more or less the same at both normal and fast rate with their small standard deviation. Secondly, the SSCF Angle12 values produced by all eight speakers are more or less the same value for the /ai/ transition. These mean SSCF Angle12 for the /ai/ sequences is an invariant characteristic and do not depend on speakers.

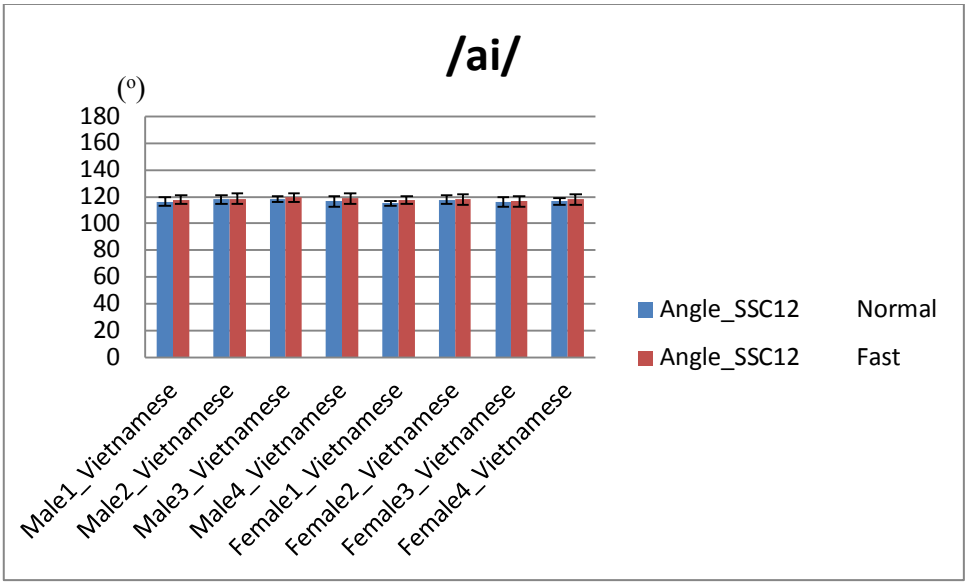


Figure 4-27: The average value and standard deviation of SSCF Angle12 of /ai/ produced by 8 Vietnamese subjects (4 males + 4 females) at both normal and fast rate.

b, SSCF Angle23 of /ai/

The average results of SSCF Angle23 of the /ai/ transition produced by eight Vietnamese native speakers (four males and four females) are presented in Figure 4-28. For each speaker, the average values of SSCF Angle23 of /ai/ transition are positive and more or less the same at both normal and fast rate. Their standard deviations are small. The SSCF Angle23 values of the /ai/ transitions produced by eight speakers are more or less the same value. Similar to SSCF Angle12, the SSCF Angle23 values of the /ai/ transitions also do not depend on speakers and is an invariant speech characteristic.

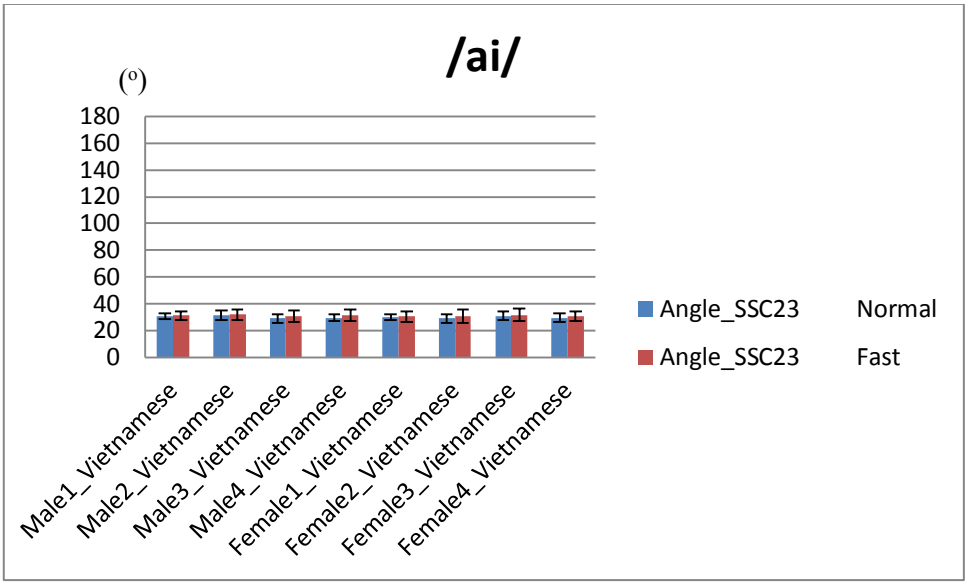


Figure 4-28: The average value and standard deviation of SSCF Angle23 of /ai/ produced by 8 Vietnamese subjects (4 males + 4 females) at both normal and fast rate.

c, SSCF Angle34 of /ai/

Figure 4-29 presents the average results and its standard deviation of SSCF Angle34 of the /ai/ transitions produced by eight Vietnamese speakers (four males and four females). For each subject, the average values of SSCF Angle34 of /ai/ transition are negative and more or less the same at both normal and fast rate. Their standard deviations are small. And the SSCF Angle34 values produced by eight speakers are more or less the same for /ai/ transition.

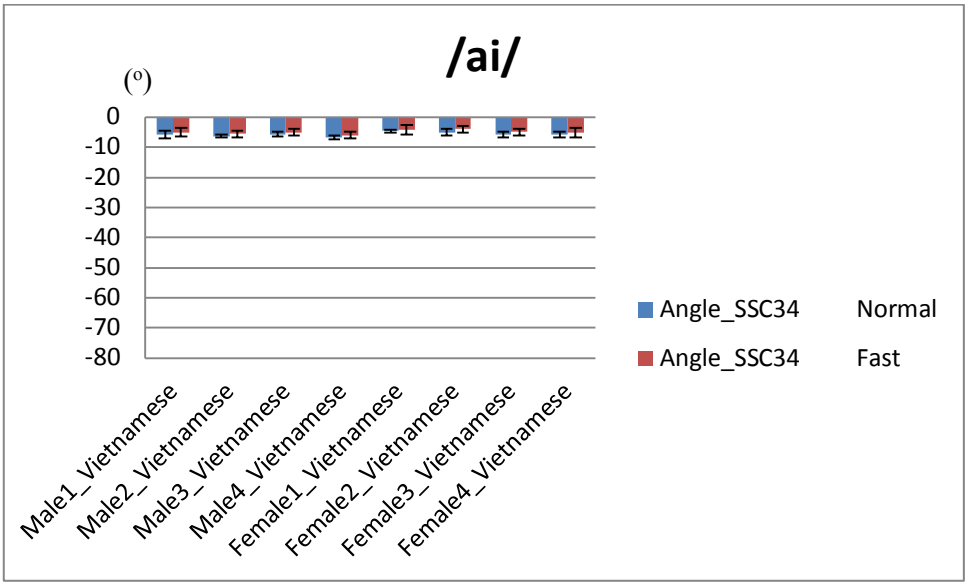


Figure 4-29: The average value and standard deviation of SSCF Angle34 of /ai/ produced by 8 Vietnamese subjects (4 males + 4 females) at both normal and fast rate.

In summary, for /ai/ transition, each SSCF Angle is more or less invariant with the different speakers.

4.4.2.2.2 /au/ sequence

a, SSCF Angle12 of /au/

Figure 4-30 presents the average results and its standard deviation of SSCF Angle12 of the /au/ transitions produced by eight Vietnamese native speakers (four males and four females). Firstly, observing each speaker, the average values of SSCF Angle12 of /au/ transitions are negative and more or less the same with their small standard deviation at both normal and fast rate. Secondly, the SSCF Angle12 values produced by all eight speakers are more or less the same value for the /au/ transition. As in the case of /ai/ transition, the SSCF Angle12 values for the /au/ sequences do not depend on speakers and they seem to be invariant among males and females.

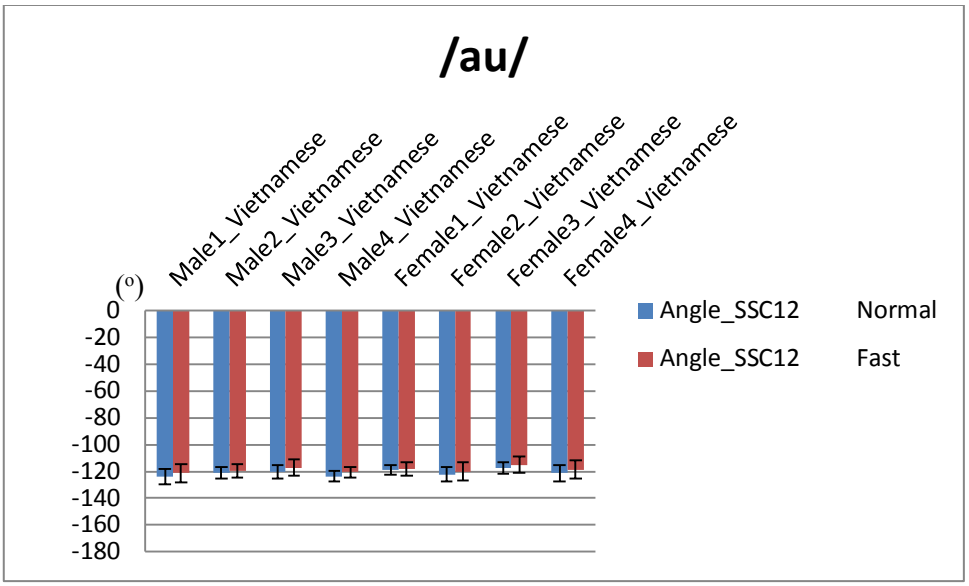


Figure 4-30: The average value and standard deviation of SSCF Angle12 of /au/ produced by 8 Vietnamese subjects (4 males + 4 females) at both normal and fast rate.

b, SSCF Angle23 of /au/

The average values of SSCF Angle23 of the /au/ transition produced by eight Vietnamese native speakers (four males and four females) were presented in Figure 4-28. For each speaker, the average values of SSCF Angle23 of /au/ transition are positive and more or less the same at both normal and fast rate. Their standard deviations are small. The SSCF Angle23 values of the /au/ transitions produced by eight speakers are more or less the same value. Similar to SSCF Angle12, the SSCF Angle23 of /au/ transitions also is independent with different speakers and is an invariant characteristic of speech.

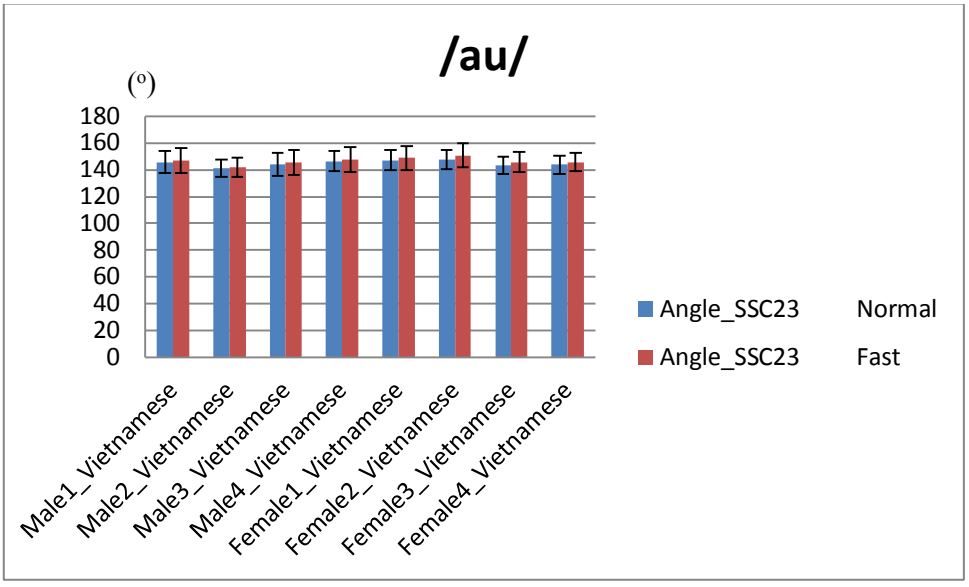


Figure 4-31: The average value and standard deviation of SSCF Angle23 of /au/ produced by 8 Vietnamese subjects (4 males + 4 females) at both normal and fast rate.

c, SSCF Angle34 of /au/

Figure 4-32 presents the average values and their standard deviation of the SSCF Angle34 of the /au/ transitions produced by eight Vietnamese native speakers (four males and four females). Firstly, although the standard deviations of the SSCF Angle34 of /au/ transitions are a bit large, but their averages are positive and more or less the same with both normal and fast rate for each speaker.

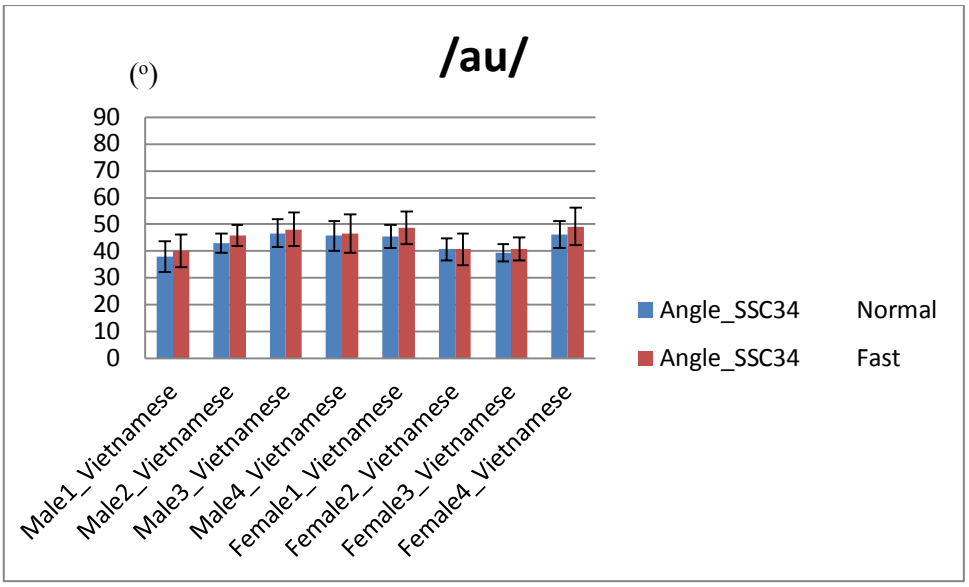


Figure 4-32: The average value and standard deviation of SSCF Angle34 of /au/ produced by 8 Vietnamese subjects (4 males + 4 females) at both normal and fast rate.

Secondly, there seem to be two groups having more or less the same values of SSCF Angle34: group 1 includes male3, male4, female1 and female4; group 2 contains male1, male2, female3 and female3. However, there are both male and female speakers in two groups, and this difference between two groups is very small. This means the SSCF Angle34 values do not discriminate with male speakers or female speakers, they only have small difference depending on specific speakers.

4.4.2.2.3 /iu/ sequence

a, SSCF Angle12 of /iu/

The average values and their standard deviations of SSCF Angle12 of the /iu/ transition produced by eight Vietnamese native speakers (four males and four females) were showed in Figure 4-33. Firstly, observing each speaker, the average values of SSCF Angle12 of /iu/ transition are negative and more or less the same with their small standard deviations at both normal and fast rate. Secondly, the SSCF Angle12 values produced by all eight speakers are more or less the same value for the /iu/ transition. The SSCF Angle12 values for the /iu/ sequences are independent with different speakers and they seem to be invariant among males and females.

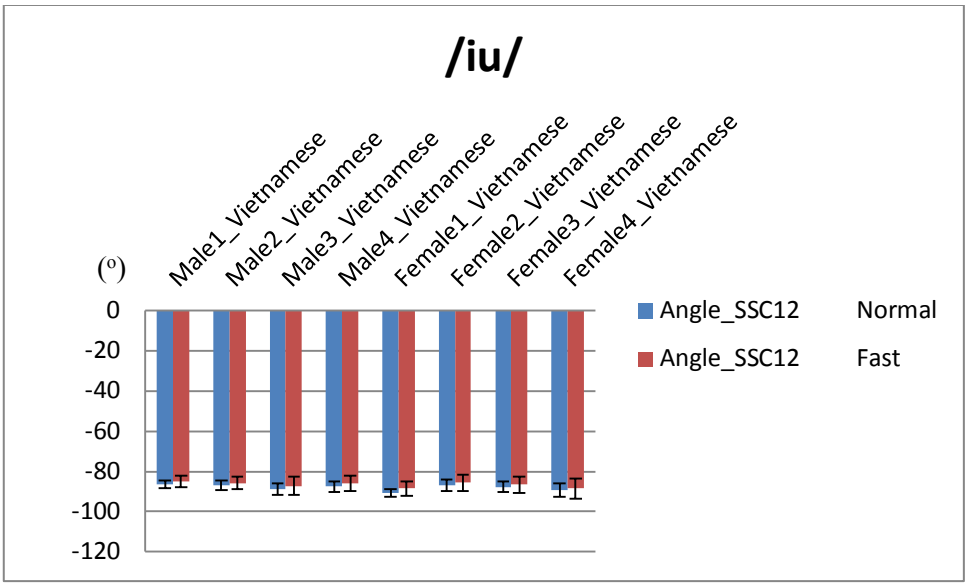


Figure 4-33: The average value and standard deviation of SSCF Angle12 of /iu/ produced by 8 Vietnamese native speakers (4 males + 4 females) at both normal and fast rate.

b, SSCF Angle23 of /iu/

Figure 4-34 presents the average values and their standard deviations of the SSCF Angle23 of the /iu/ transitions produced by eight Vietnamese native speakers (four males and four females). The averages of SSCF Angle23 of /iu/ transitions are negative and almost -180^0, they are more or less the same values at both normal and fast rate. Their standard deviations are very small. Moreover, the SSCF Angle23 values produced by all eight speakers are more or less the same for /iu/ transitions.

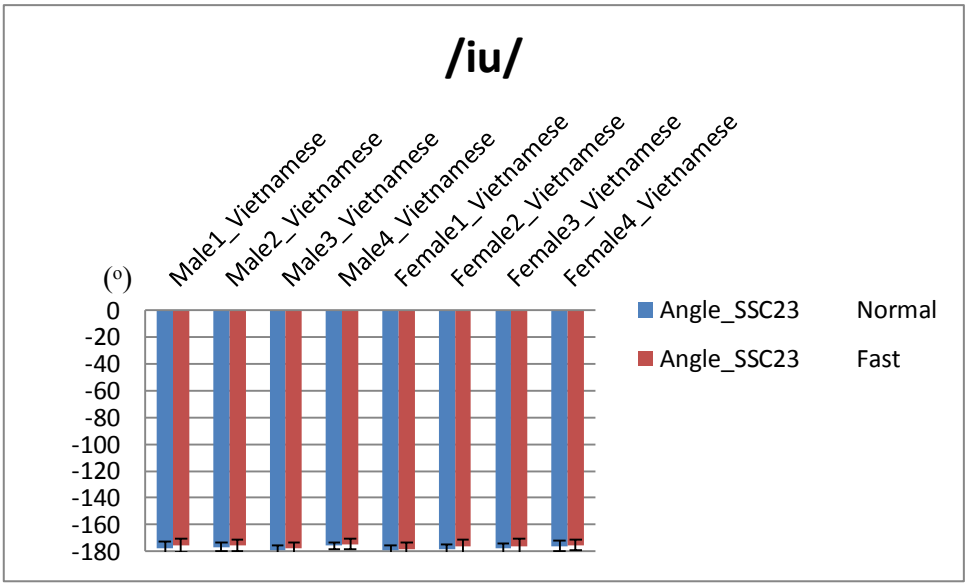


Figure 4-34: The average value and standard deviation of SSCF Angle23 of /iu/ produced by 8 Vietnamese native speakers (4 males + 4 females) at both normal and fast rate.

c, SSCF Angle34 of /iu/

Figure 4-35 presents the average values and their standard deviations of the SSCF Angle34 of the transition /iu/ produced by eight Vietnamese native speakers (four males and four females). We found out that the averages of SSCF Angle34 of /iu/ transition are positive and more or less the same with their very small standard deviation at both normal and fast rate. And they are more or less the same values among eight speakers. Obviously, the SSCF Angle34 parameters for /iu/ transitions are independent with different speakers and are invariant among male and female speakers.

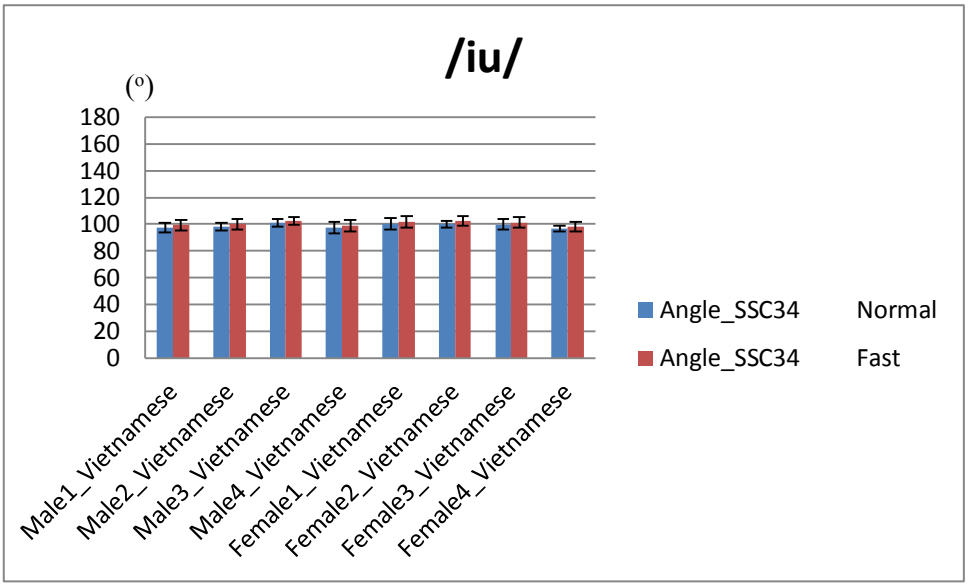


Figure 4-35: The average value and standard deviation of SSCF Angle34 of /iu/ produced by 8 Vietnamese native speakers (4 males + 4 females) at both normal and fast rate.

4.4.2.2.4 Other Vietnamese V1V2 transition sequences

Following up the three SSCF Angles of the three transition sequences /ai, au, iu/ which are three boundaries of acoustic vowel triangle, the two same results are obtained. Firstly, for each transition sequence, the averages of each SSCF Angle are more or less the same with their very small standard deviation at both normal and fast rate. Secondly, the values for each SSCF Angle for each transition produced by all eight speakers are more or less the same. This means that the values of each SSCF Angle for each transition sequence are independent with different speakers and are invariant among male and female speakers.

These characteristics are also observed for other transition sequences /aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, aɔ, ao/ presented in Appendix 2.

4.4.2.3 Vietnamese V1V2 transitions in 3-D plane of SSCF Angles

Our first results on SSCF Angles comparisons among fourteen transitions for each speaker were studied in section 4.4.2.1. SSCF Angles comparisons with the same transitions produced by eight Vietnamese native speakers (4 males + 4 females) also presented in section 4.4.2.2. In this section, we

compare the different VV transitions produced by the same eight Vietnamese native speakers by the combination of three SSCF Angles (SSCF Angle12, SSCF Angle23 and SSCF Angle34) in 3-D plane.

Following up the results of two cases in the previous sections, we found that for each VV transition, the values of each SSCF Angle are more or less the same at both normal and fast rate for each speaker. Therefore, in this section, we only present the results of different VV transitions at normal rate for each speaker.

Keeping up the results of case 1, six groups from the fourteen transitions /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ are completely separated if based on SSCF Angle12 or SSCF Angle23: group 1 includes the transitions of /ai, aɛ, ae/; group 2 consist of /ui/ transitions; group 3 contains /ua, ɔa, oa/ transitions; group 4 involves /ia, ɛa, ea/ transitions; group 5 is the /iu/ transitions; and group 6 consists of the transition of /au, aɔ, ao/. However, all these fourteen VV transitions were completely discriminated based on their SSCF Angle34. On this basis, if each transition in four groups 1, 3, 4 or 6 can be separated mutually, then all fourteen transitions /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ will be completely separated. This point will be considered as in following presentation.

4.4.2.3.1 Group of /ai, aɛ, ae/ transitions in 3-D plane of SSCF Angles

Figure 4-36 shows the /ai, aɛ, ae/ transitions produced by the same eight Vietnamese native speakers in 3-D plane which the x-, y-, and z-axis are SSCF Angle12, SSCF Angle34 and SSCF Angle23, respectively.

In this figure, the transitions of each speaker were displayed the same color, but with different markers, such as plus points were defined for /ai/ transitions; circle points were assigned for /aɛ/ transitions; and square points were defined for /ae/ transitions.

Observing Figure 4-36 from this view direction, we found that firstly, the /ai/ transitions produced by eight Vietnamese native speakers (4 males + 4 females) converge in one region in this 3-D space. This corresponding result is also obtained with the /aɛ/ or /ae/ transitions. Secondly, three converging regions correspond to /ai/, /aɛ/ and /ae/ transitions are completely separated mutually in the 3-D plane of SSCF Angles.

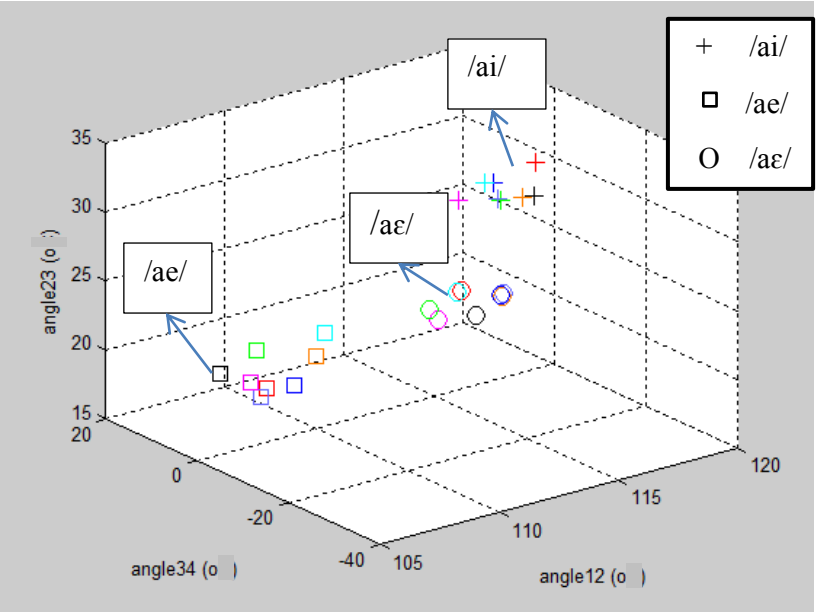


Figure 4-36: /ai, aɛ, ae/ transitions produced 8 Vietnamese native speakers (4 males + 4 females) in 3-D plane of (SSCF Angle12, SSCF Angle34 and SSCF Angle23).

4.4.2.3.2 Group of /ia, ɛa, ea/ transitions in 3-D plane of SSCF Angles

Do the same way with the group 1, we continue to plot /ai, aɛ, ae/ transitions in 3-D angle space which the SSCF Angle12, SSCF Angle34 and SSCF angl23 are corresponding to x-, y- and z-axis, respectively.

In this figure, the transitions of each speaker were displayed the same color, but with different markers, such as diamond points were defined for /ea/ transitions; cross points were assigned for /ɛa/ transitions; and star points were defined for /ia/ transitions.

From this view, the /ia/, /ɛa/ and /ea/ transitions converge in three distinct regions in 3-D plane of SSCF Angles, respectively. And these three regions are completely separated mutually.

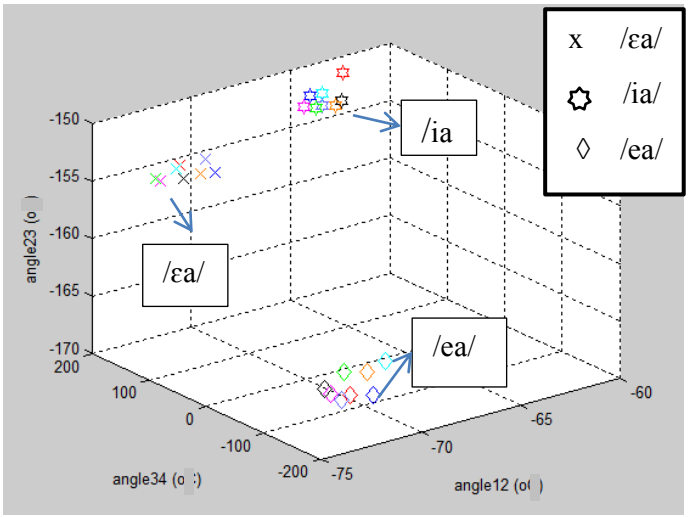


Figure 4-37: /ia, ɛa, ea/ transitions produced 8 Vietnamese native speakers (4 males + 4 females) in 3-D plane of (SSCF Angle12, SSCF Angle34 and SSCF Angle23).

4.4.2.3.3 Group of /oa, ɔa, ua/ in 3-D plane of SSCF Angles

Figure 4-36 shows the /oa, ɔɛ, ua/ transitions produced by the same eight Vietnamese native speakers in 3-D plane which the x-, y-, and z-axis are SSCF Angle12, SSCF Angle34 and SSCF Angle23, respectively. These transitions of each speaker were displayed the same color, but with different markers, such as diamond points were defined for /oa/ transitions; cross points were assigned for /ɔa/ transitions; and star points were defined for /ua/ transitions.

Three distinct regions correspond to the transitions of /oa/, /ɔa/ and /ua/ produced by eight Vietnamese native speakers, are pointed out in Figure 4-38. These regions are well separated.

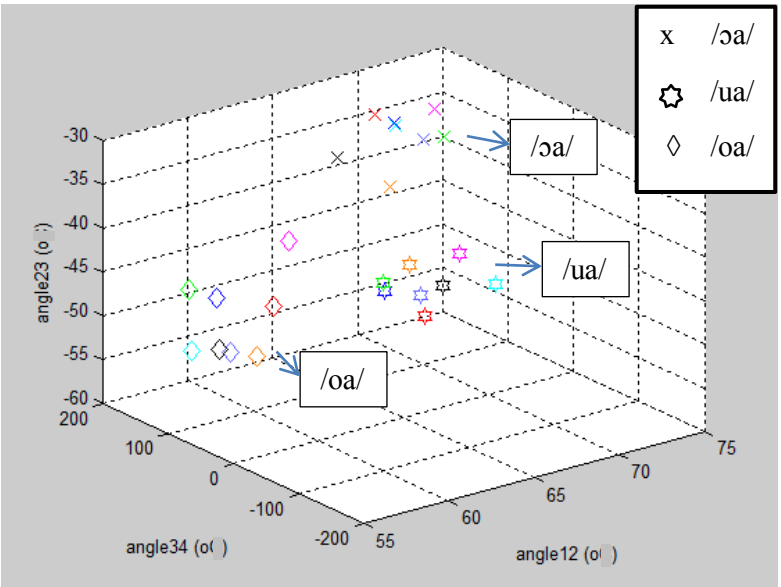


Figure 4-38: /oa, ɔa, ua/ transitions produced 8 Vietnamese native speakers (4 males + 4 females) in 3-D plane of (SSCF Angle12, SSCF Angle34 and SSCF Angle23).

4.4.2.3.4 Group of /ao, aɔ, au/ in 3-D plane of SSCF Angles

Performing the same way with the above groups, we present /ao, aɔ, au/ transitions in 3-D angle space which x-, y- and z-axis are the SSCF Angle12, SSCF Angle34 and SSCF angl23, respectively. In this figure, these /ao, aɔ, au/ transitions produced by each speaker, were displayed with the same color, but with different markers, such as square points were defined for /ao/ transitions; circle points were assigned for /aɔ/ transitions; and plus points were defined for /au/ transitions.

Observing Figure 4-39, we found that firstly, the /au/ transitions produced by eight Vietnamese native speakers (4 males + 4 females) converge in one region in this 3-D space. This corresponding result is also obtained with the /ao/ or /aɔ/ transitions. Secondly, these three converging regions corresponding to /ao/, /aɔ/ and /au/ transitions are completely separated mutually in the 3-D plane of SSCF Angles.

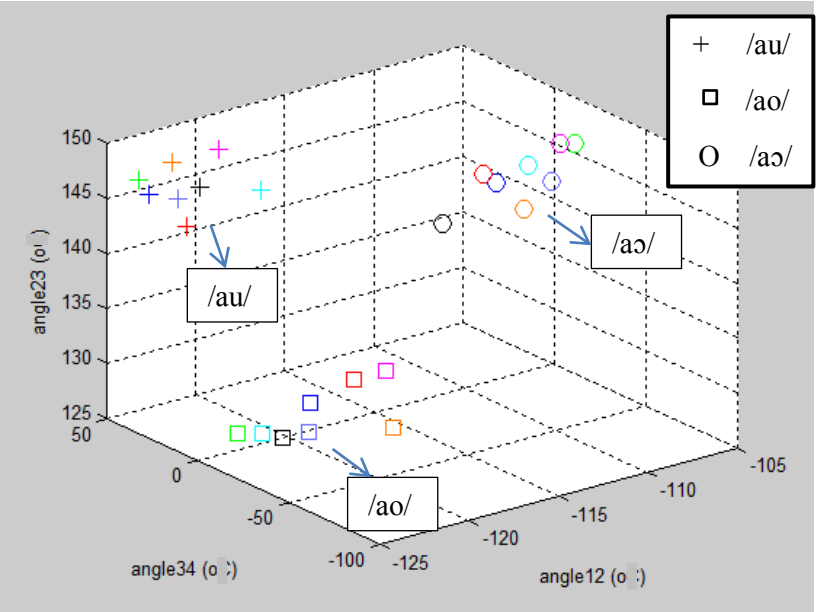


Figure 4-39: /oa, ɔa, ua/ transitions produced 8 Vietnamese native speakers (4 males + 4 females) in 3-D plane of (SSCF Angle12, SSCF Angle34 and SSCF Angle23).

Following up all the obtained results of Group 1, 3, 4, and 6 in 3-D plane of (SSCF Angle12, SSCF Angle34 and SSCF Angle23), (i) firstly, we pointed out that the transitions of the same V1V2 produced by eight Vietnamese speakers in each group converged in one distinct region, and (ii) the regions corresponding to the different V1V2 transitions in each group were completely separated, therefore (iii), the fourteen transitions /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produced by eight Vietnamese native speakers converged in the corresponding fourteen regions in 3-D space of SSCF Angles and these fourteen regions are completely distinguished from one another.

4.4.3 Conclusions

Fourteen transitions of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ were compared in two cases: (i) first case is the SSCF Angles comparisons among different items for each speaker in section 4.4.2.1; (ii) second case is the SSCF Angles comparisons with the same item produced by eight Vietnamese native speakers (4 males + 4 females). And finally, all fourteen transitions were presented in 3-D plane of SSCF Angles in section 4.4.2.3. Some interesting results were obtained: (i) for each speaker, each SSCF Angle is more or less invariant at both normal and fast speech rate for each V1V2 transition; (ii) for each speaker, all fourteen transitions can be completely separated mutually in 3-D plane of SSCF Angles; (iii) for the same V1V2 transitions, each SSCF Angle12, SSCF Angle23 or SSCF Angle34 is more or less the same value for eight studied Vietnamese native speakers which included 4 males and 4 females, in other words, each SSCF Angle do not depend on different speakers; (iv) the transitions of the same V1V2 transitions produced by eight Vietnamese speakers converged in one regions in 3-D space of SSCF Angles, and therefore, there were fourteen distinct regions correspond to fourteen studied transitions based on the combination of three SSCF Angle12,

SSCF Angle34 and SSCF Angle23 in 3-D plane, in other words, the fourteen transitions of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produced by eight Vietnamese speakers were completed separated mutually; (v) besides, V1V2 and V2V1 transitions had symmetrical properties on the acoustic domain based on SSCF Angles because the sum of the absolute value of their angles is roughly equal to 180^o.

4.5 SSCF Angles analysis on French Vowel-to-Vowel transitions

In section 4.4, we performed SSCF Angles analysis in Vietnamese V1V2 transitions. The obtained results showed that we can separate completely these fourteen transitions of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ based on the combination of three SSCF Angles (SSCF Angle12, SSCF Angle23 and SSCF Angle34). This means that SSCF Angles are useful characteristics for separating the V1V2 transitions in Vietnamese.

A question is raised here: whether SSCF Angles still work well in other languages? In order to answer a part of this question, we continue to analyze SSCF Angles on some V1V2 transitions in French in the following sections.

4.5.1 Methodology

4.5.1.1 French stimuli

In this study , we only study on the five transitions of /ai, aɛ, ae, ua, ui/ in French in order to compare with the obtained results in Vietnamese.

This very small French corpus was built by four French native speakers (two males and two females) with ages from 25 to 30 years old.

Each V1V2 sequence was inserted in the carrier sentence: “Dites V1V2 trois fois” (“Speak V1V2 three times”). And as in Vietnamese stimuli, each French V1V2 sequence was produced ten times at two speech rate (normal and fast rate). Table 4-2 shows in detail the total studied sentence number in Vietnamese. There were 200 recording sentences for four French speakers.

The speaker was instructed to read the V1V2 sequences so that the transition from V1 vowel to V2 vowel is continuous in time domain. The recording process was controlled by PC software that randomly presented the succession of the items to be recorded. In case of bad pronunciation or hesitation, the speaker had to pronounce the item again.

This recording took place in quiet studio of GIPSA-lab, Grenoble, France. The corpus was recorded by microphone with a sampling frequency of 16.000Hz and 16bits per sample and save in .wav format.

The speaker was invited to read the corpus at normal speed. A small break was made during reading the different sentences. After completing the corpus at normal speed, the speaker was invited to read the corpus at fast speed. This process was going on repetitively for ten times of the recording.

Table 4-2: Total number of the studied sentences in French.

	Number of items/speaker	Recording times/item/sentence		Total studied sentence number
		At normal rate	At fast rate	
1 speaker	5	5	5	50 sentences/speaker
4 speakers	5	5	5	200 sentences/4 speakers

4.5.1.2 Analysis method

Performing the same method in Vietnamese analysis, we also compare the three SSCF Angles (SSCF Angle12, SSCF Angle23 and SSCF Angles34) in French V1V2 transitions according to two cases, as follows:

- first case is the SSCF Angles comparisons among different items for each speaker;
- second case is the SSCF Angles comparisons with the same item among males and females.

In each case, the SSCF Angles are also compared at between normal and fast speech rate for same item and same speaker.

4.5.2 Results

Each SSCF Angle at each speech rate (normal/fast) is presented by the angle average and its standard deviations. Each item was produced five times at normal rate and five times at fast rate. The analysis results are presented in the following sections.

4.5.2.1 Case 1: SSCF Angles comparisons among different transitions for each speaker

As in Vietnamese analysis, we will present the typical results with one French male M1_Fr and one French female F1_Fr. The other results of other French males and females will be presented in Appendix 3.

4.5.2.1.1 SSCF Angle12

Figures 4-40 and 4-41 present the average and standard deviation of SSCF Angle12 of the five transitions of /ai, aɛ, ae, ua, ui/ produced by one French native male (M1_Fr) and one French native female (F1_Fr), respectively.

Figure 4-40 shows the result of SSCF Angles of one French male participant M1_Fr. Firstly, for each V1V2 transition sequence, the averages of SSCF Angle12 are more or less the same values and its standard deviation is small at both normal and fast rate. This indicates the invariant property of SSCF Angle12 of the same VV sequences produced by the same speaker at the different rate.

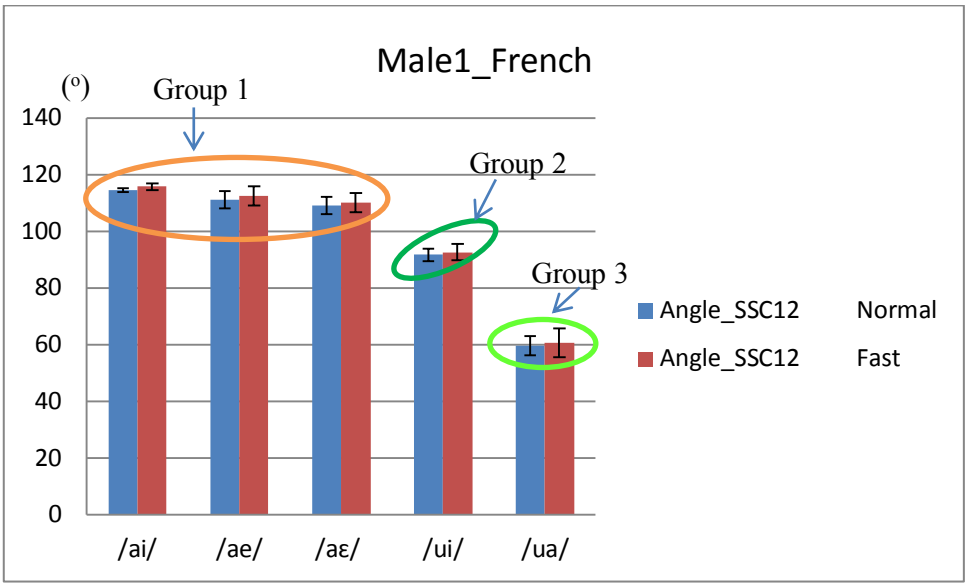


Figure 4-40: The average value and standard deviation of SSCF Angle12 of /ai, aɛ, ae, ua, ui / produced by one French male (M1_Fr) at both normal and fast rate.

Secondly, three groups of V1V2 sequences are separated according to their different SSCF Angle12 values: group 1 includes the transitions of /ai, aɛ, ae/; group 2 is the /ui/ transitions; group 3 contains /ua/ transitions. The SSCF Angle12 values of the all three groups of V1V2 are positive. These results were obtained for one French female F1_Fr in Figure 4-41. These characteristics are also observed for the other French speakers in Appendix 3.

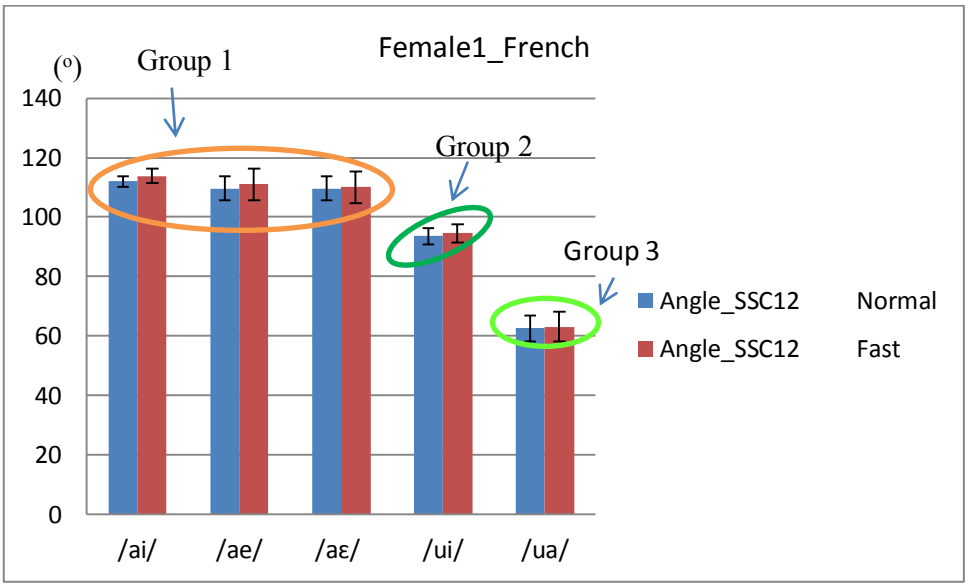


Figure 4-41: The average value and standard deviation of SSCF Angle12 of /ai, aɛ, ae, ua, ui/ produced by one French male (F1_Fr) at both normal and fast rate.

4.5.2.1.2 SSCF Angle23

The average and standard deviation values of SSCF Angle23 of the five transitions /ai, aɛ, ae, ua, ui/ produced by the same one male M1_Fr and one female F1_Fr are presented in Figures 4-42 and 4-43, respectively.

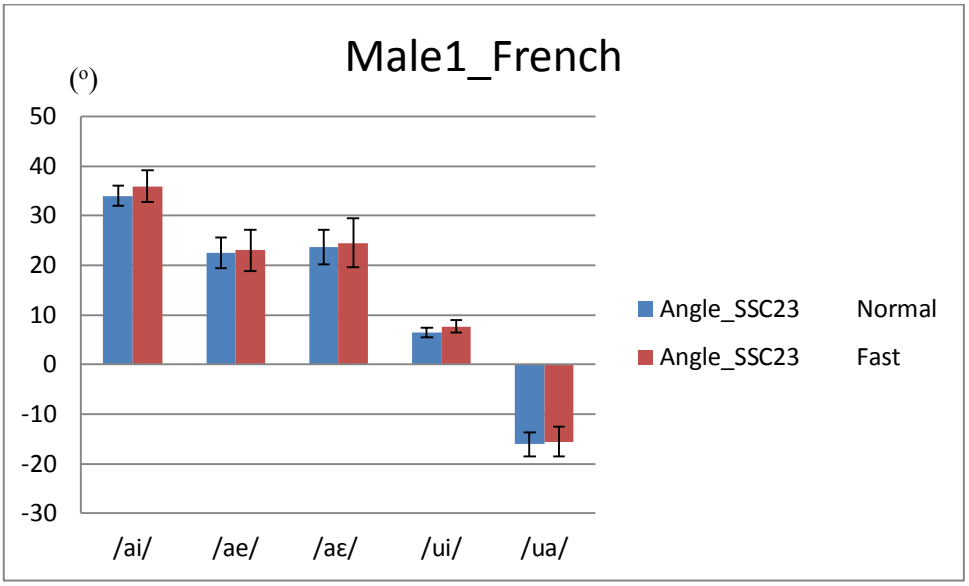


Figure 4-42: The average value and standard deviation of SSCF angl23 of /ai, aɛ, ae, ua, ui / produced by one French male (M1_Fr) at both normal and fast rate.

Some similar results are obtained for both French male M1_Fr and French female F1_Fr. Firstly, for each item (/ai/, /aɛ/, /ae/, /ui/ or /ua/), although the standard deviations are a bit large, but their SSCF Angle23 values are more or less the same at both normal and fast rate. Secondly, while the values of SSCF Angle23 are positive for /ai, aɛ, ae, ui/ transitions, the one is negative for /ua/ transitions. Thirdly, the SSCF Angle23 values of /ae/ and /aɛ/ transitions are fairly similar, but the SSCF Angle23 are different among this group of /ae, aɛ/ and the remaining transitions of /ai, ui, ua/.

These results are also observed for other French speakers in Appendix 3.

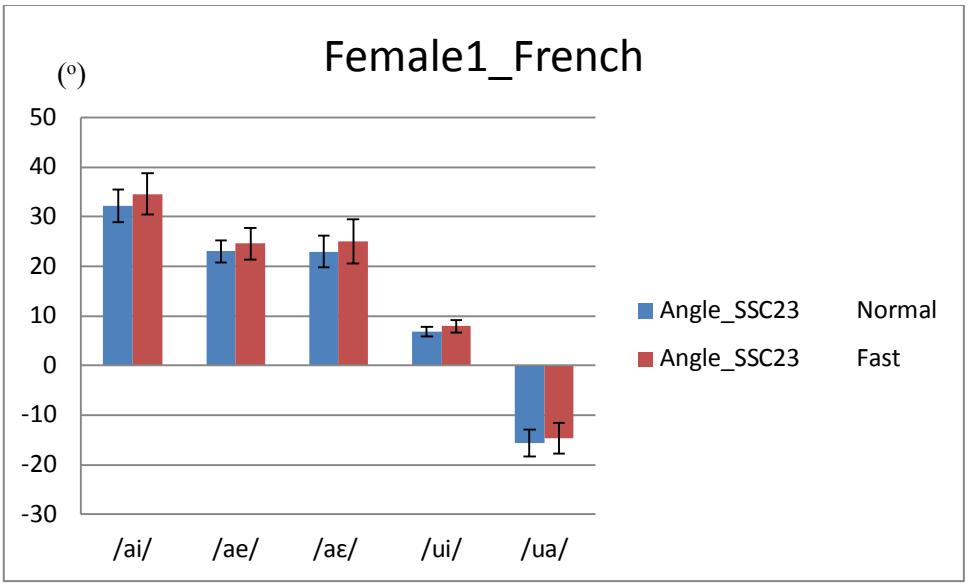


Figure 4-43: The average value and standard deviation of SSCF Angle23 of /ai, aɛ, ae, ua, ui/ produced by one French male (F1_Fr) at both normal and fast rate.

4.5.2.1.3 SSCF Angle34

Figures 4-44 and 4-45 present the average and standard deviation of SSCF Angle34 for the five transitions /ai, aɛ, ae, ua, ui/ produced by the same male speaker M1_Fr and female speaker F1_Fr, respectively.

For both French male M1_Fr and French female F1_Fr, we observe the similar results. Firstly, for each item (/ai/, /aɛ/, /ae/, /ua/ or /ui/) of each speaker, the averages of SSCF Angle34 are more or less the same values at both normal and fast rate with its small standard deviation. Secondly, while the SSCF Angle34 values are negative for /ai, aɛ, ui, ua/, the one is positive for /ae/ transitions. Thirdly, the SSCF Angle34 values are completely different for these five transitions of /ai, aɛ, ae, ua, ui/.

Obviously, these five French transitions of /ai, aɛ, ae, ua, ui/ were completely separated according to their SSCF Angle34.

These same results on SSCF Angle34 are also observed for other French speakers in Appendix 3.

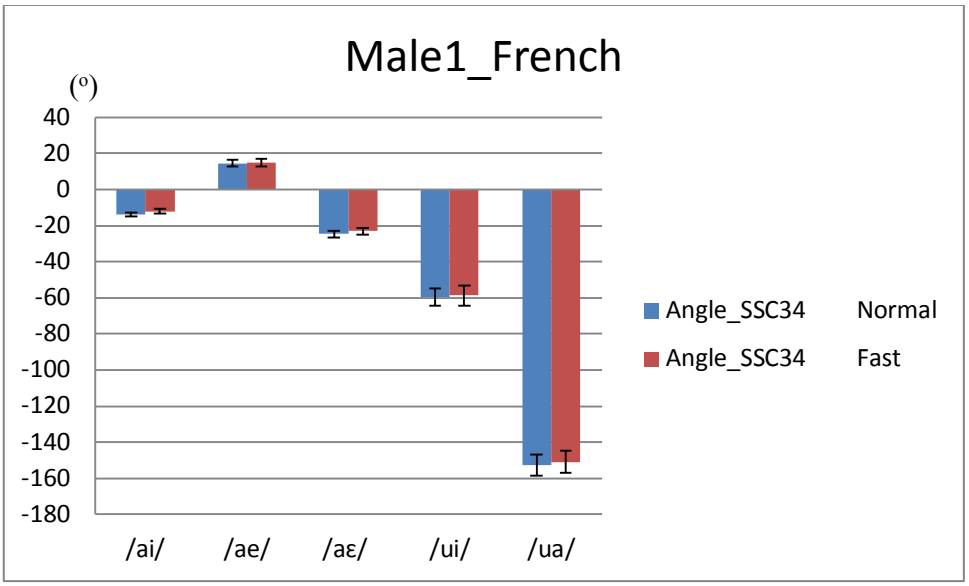


Figure 4-44: The average value and standard deviation of SSCF Angle34 of /ai, aɛ, ae, ua, ui/ produced by one French male (M1_Fr) at both normal and fast rate..

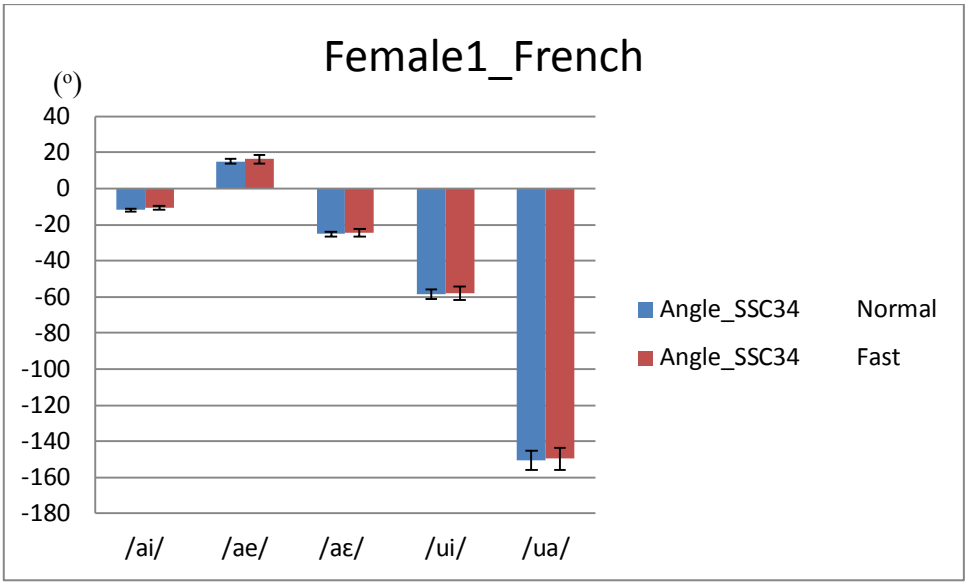


Figure 4-45: The average value and standard deviation of SSCF Angle23 of /ai, aɛ, ae, ua, ui/ produced by one French male (F1_Fr) at both normal and fast rate.

In summary, in this case of SSCF Angles comparisons among different transitions for each speaker, following up the obtained results of SSCF Angle12, SSCF Angle23 and SSCF Angle34, there is similar result between normal and fast production for each SSCF Angle in each transition sequence produced by each speaker. And while these five French transitions of /ai, aɛ, ae, ua, ui/ were not completely separated if based on SSCF Angle12 or SSCF Angle23, but these five transitions can be completely distinguished based on their SSCF Angle34.

4.5.2.2 Case 2: SSCF Angles comparisons with the same transition among males and females

Our French results on case 1 were presented the differences of five French VV productions produced by each French speaker. In this case 2, we consider each VV transition produced by four French native speakers.

4.5.2.2.1 /ai/ sequence

a, SSCF Angle12 of /ai/

Figure 4-46 presents the average and its standard deviation of SSCF Angle12 of the /ai/ transition produced by four French speakers (2 males and 2 females). We found that firstly, for each speaker, the average values of SSCF Angle12 of /ai/ transition are positive and more or less the same at both normal and fast rate with their small standard deviation. Secondly, the SSCF Angle12 values produced by all four Vietnamese speakers are more or less the same value for the /ai/ transitions. These mean SSCF Angle12 for the /ai/ sequence is an invariant characteristic and do not depend on speakers.

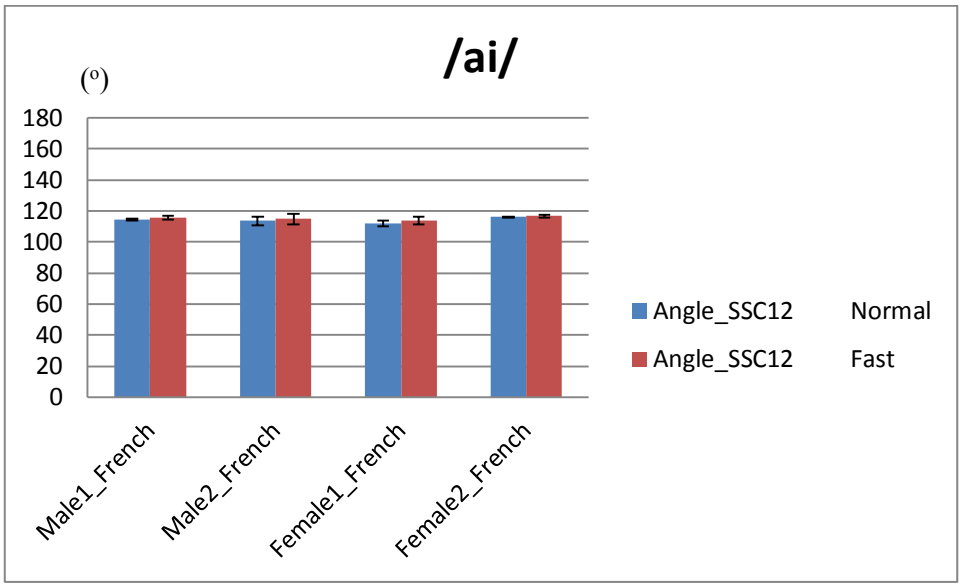


Figure 4-46: The average value and standard deviation of SSCF Angle12 of /ai/ produced by 4 French subjects (2 males + 2 females) at both normal and fast rate.

b, SSCF Angle23 of /ai/

The average results of SSCF Angle23 of the /ai/ transition produced by four French native speakers (2 males and 2 females) were presented in Figure 4-47. Firstly, although the standard deviations are not very small, but as a whole, the SSCF Angle23 values are positive and seem to be more or less the same at both normal and fast rate for /ai/ transition for each speaker. Secondly, the SSCF Angle23 values of /ai/ transition produced by four French speakers are more or less the same. Similar to SSCF Angle12, the SSCF Angle23 of /ai/ transition is also independent with speakers and it is an invariant speech characteristic.

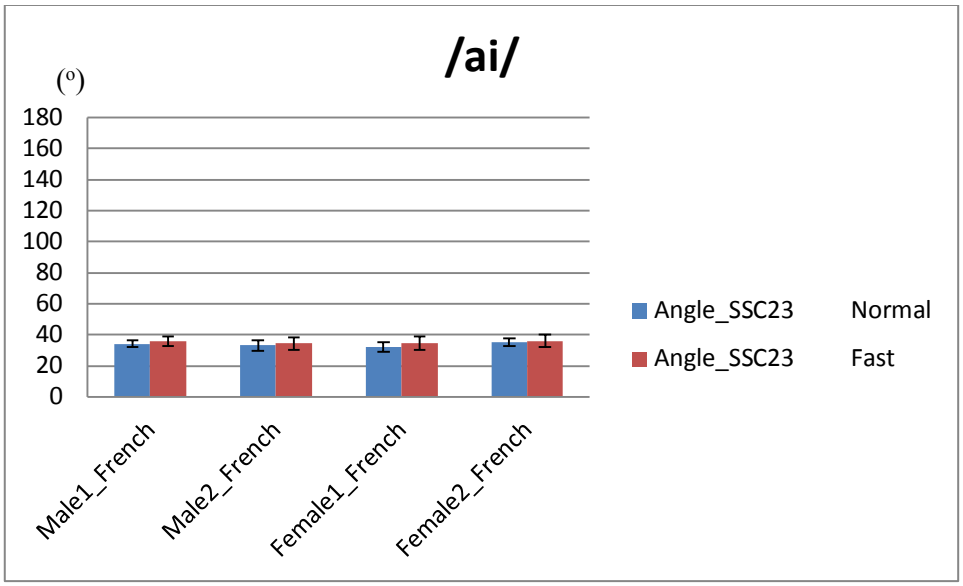


Figure 4-47: The average value and standard deviation of SSCF Angle23 of /ai/ produced by 4 French subjects (2 males + 2 females) at both normal and fast rate.

c, SSCF Angle34 of /ai/

Figure 4-49 presents the averages and their standard deviation of SSCF Angle34 of the /ai/ transition produced by four French speakers (2 males and 2 females). We realized that firstly, the average values of SSCF Angle34 of /ai/ transition are negative and more or less the same at both normal and fast rate for each subject. Their standard deviations are small. Secondly, the SSCF Angle34 values produced by four French speakers are more or less the same for /ai/ transitions.

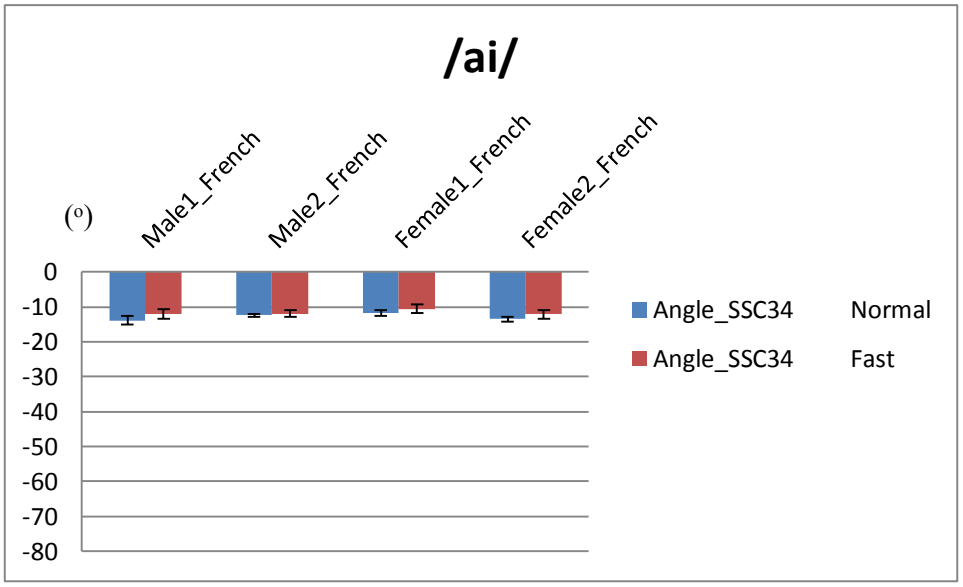


Figure 4-48: The average value and standard deviation of SSCF Angle34 of /ai/ produced by 4 French subjects (2 males + 2 females) at both normal and fast rate.

In summary, each SSCF Angle is fairly invariant with the different speakers for /ai/ transitions.

4.5.2.2.2 /ua/ sequence

a, SSCF Angle12 of /ua/

Figure 4-49 presents the average results and its standard deviation of SSCF Angle12 of the /ua/ transitions produced by four French native speakers (2 males and 2 females). Firstly, although the standard deviations are a little large, but the values of SSCF Angle12 of /ua/ transition are more or less the same at both normal and fast rate for each speaker and all are negative value. Secondly, the SSCF Angle12 values produced by all four French speakers are more or less the same value for the /ua/ transitions. As in the case of /ai/ transition, the SSCF Angle12 values for the /ua/ sequences are independent with speakers and they are invariant among males and females.

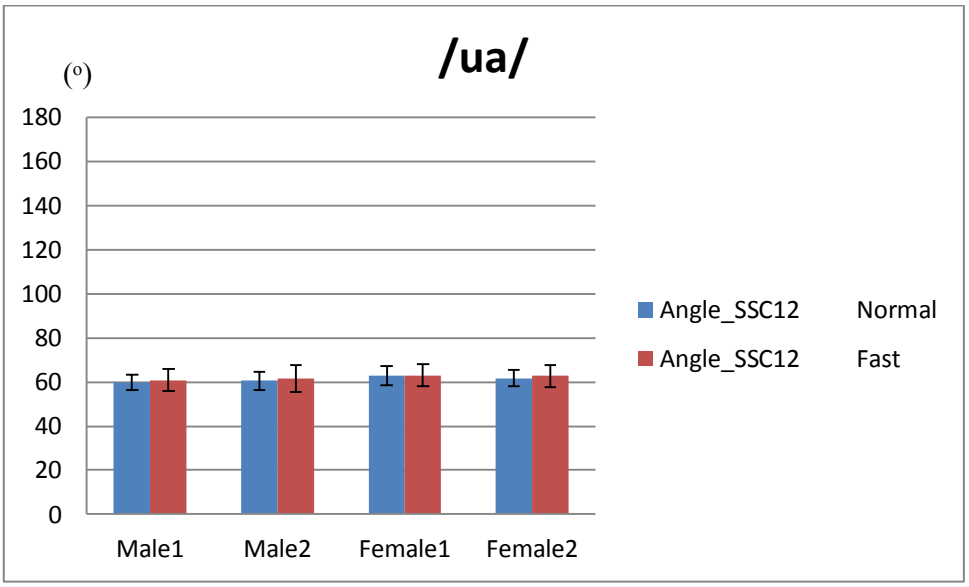


Figure 4-49: The average value and standard deviation of SSCF Angle12 of /ua/ produced by 4 French subjects (2 males + 2 females) at both normal and fast rate.

b, SSCF Angle23 of /ua/

The average values of SSCF Angle23 of the /ua/ transition produced by four French native speakers (2 males and 2 females) were presented in Figure 4-50. Observing this figure, firstly although the standard deviations are not small, but as a whole the values of SSCF Angle23 of /ua/ transition are negative and more or less the same at both normal and fast rate for each speaker.

Secondly, the SSCF Angle23 of /ua/ transitions produced by four French speakers are also more or less the same values.

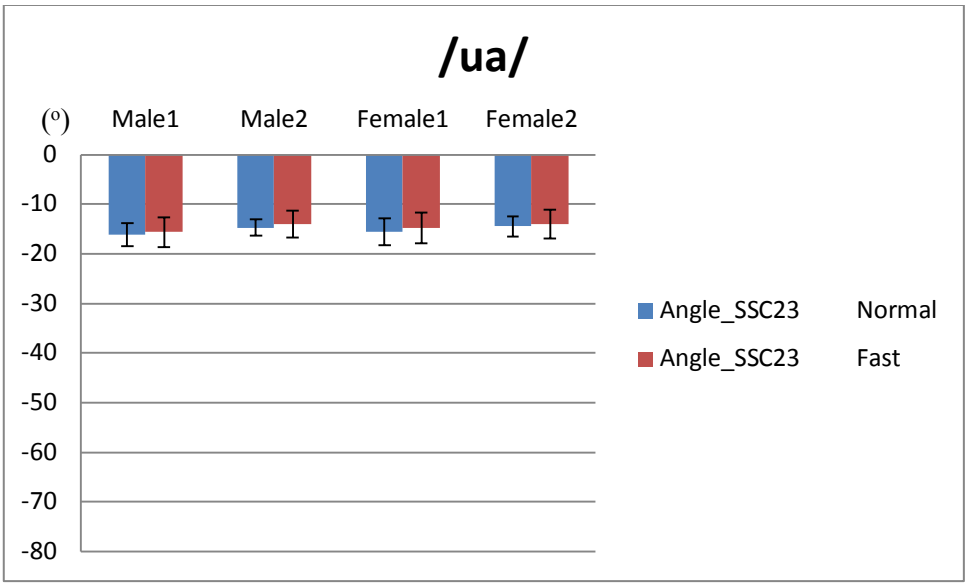


Figure 4-50: The average value and standard deviation of SSCF Angle23 of /ua/ produced by 4 French subjects (2 males + 2 females) at both normal and fast rate.

c, SSCF Angle34 of /ua/

Figure 4-51 presents the average values and their standard deviation of the SSCF Angle34 of the /ua/ transitions produced by four French native speakers (2 males and 2 females). Some comments were pointed out. Firstly, although the standard deviations of the SSCF Angle34 of /ua/ transitions are not small, but all their averages are negative and more or less the same with both normal and fast rate for each speaker.

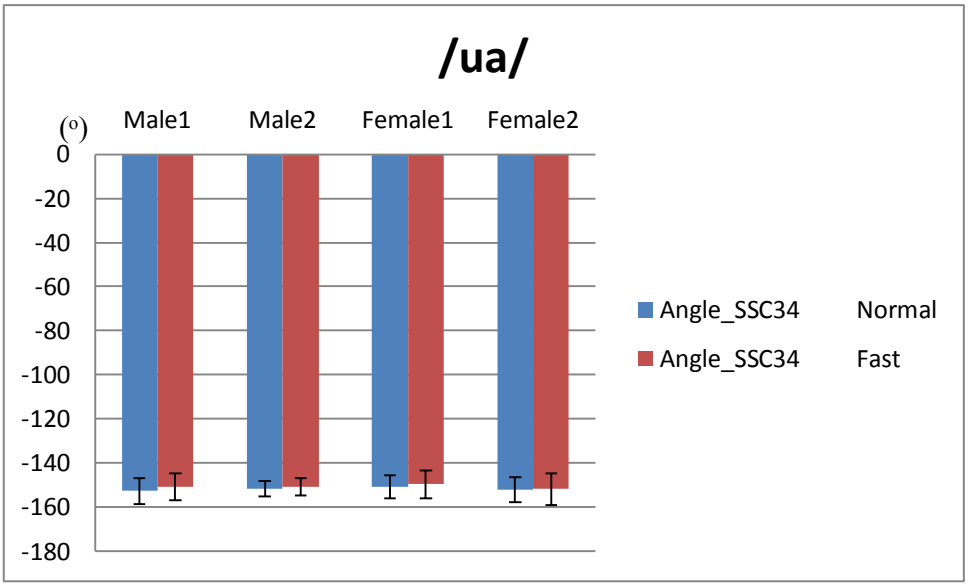


Figure 4-51: The average value and standard deviation of SSCF Angle23 of /ua/ produced by 4 French subjects (2 males + 2 females) at both normal and fast rate.

Secondly, the SSCF Angle34 of /ua/ transitions produced by four French speakers look to be fairly similar mutually. This means SSCF Angle34 seem to be the invariant speech characteristics and independent with speakers.

4.5.2.2.3 /ui/ sequence

a, SSCF Angle12 of /ui/

The average values and their standard deviations of SSCF Angle12 of the /ui/ transitions produced by four French native speakers (2 males and 2 females) were showed in Figure 4-52. Observing this figure, we found that firstly, the average values of SSCF Angle12 of /ui/ transition are positive and more or less the same at both normal and fast rate with their small standard deviations.

Secondly, the SSCF Angle12 values produced by all four French speakers are more or less the same value for the /ui/ transition. It means that the SSCF Angle12 values for the /ui/ sequences are independent with different speakers and they seem to be invariant among males and females.

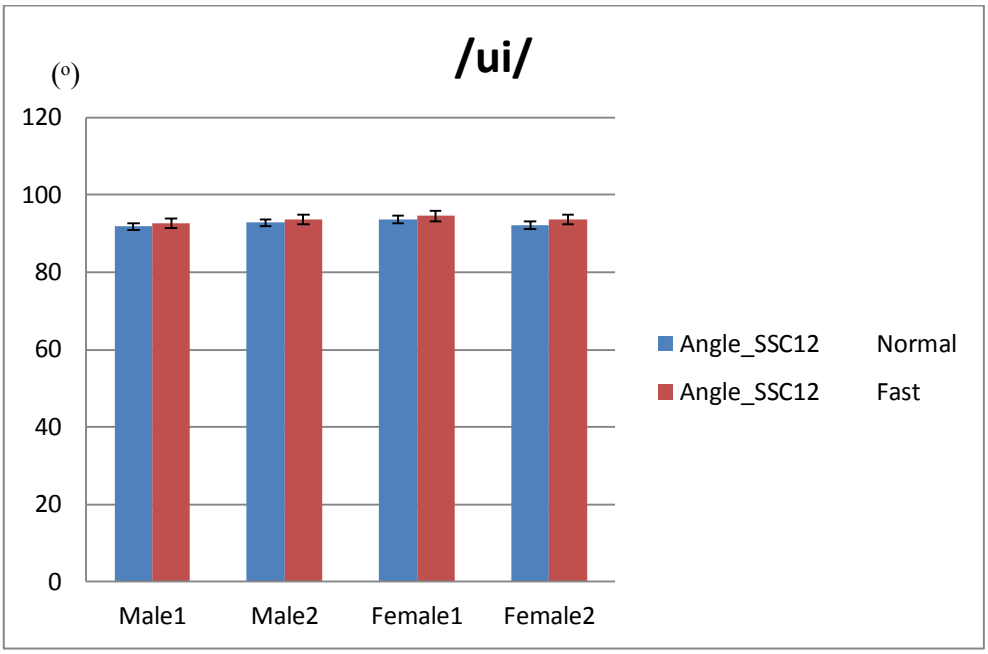


Figure 4-52: The average value and standard deviation of SSCF Angle12 of /ui/ produced by 4 French subjects (2 males + 2 females) at both normal and fast rate.

b, SSCF Angle23 of /ui/

Figure 4-53 presents the average values and their standard deviations of the SSCF Angle23 of the transition /ui/ produced by four French native speakers (2 males and 2 females). The averages of SSCF Angle23 of transition /ui/ are positive and they are more or less the same values at both normal and fast rate with the standard deviations are small for each speaker.

Moreover, the SSCF Angle23 values produced by all four French speakers are more or less the same for /iu/ transition. It means that SSCF Angle23 is invariant among males and females.

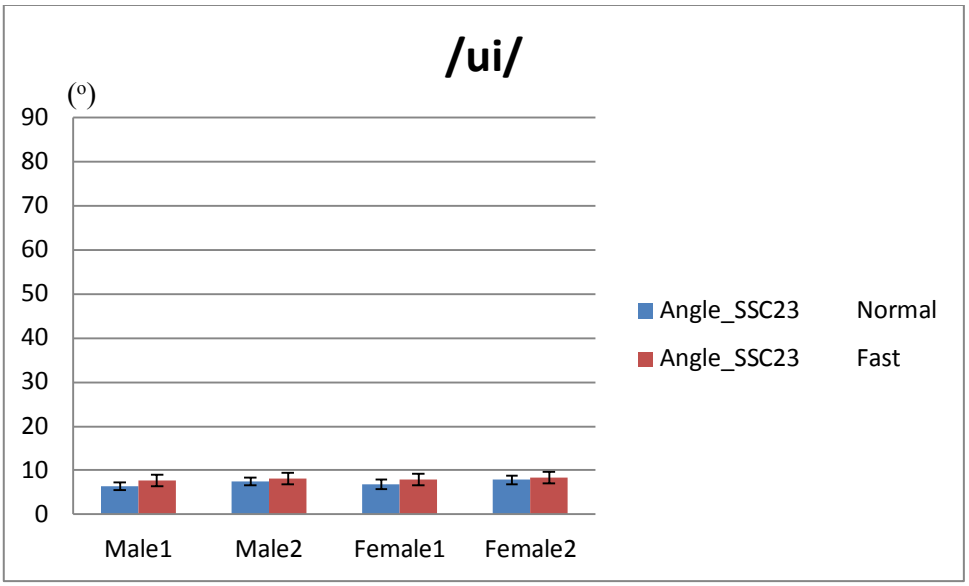


Figure 4-53: The average value and standard deviation of SSCF Angle23 of /ui/ produced by 4 French subjects (2 males + 2 females) at both normal and fast rate.

c, SSCF Angle34 of /ui/

Figure 4-54 presents the average values and their standard deviations of the SSCF Angle34 of the transition /ui/ produced by four French native speakers (2 males and 2 females). We found out that although the standard deviations are a little large, but the averages of SSCF Angle34 of /ui/ transition are negative and more or less the same at both normal and fast rate. And they are also more or less the same values among four French speakers. Obviously, the SSCF Angle34 parameters for /ui/ transitions are independent with different speakers and are invariant among male and female speakers.

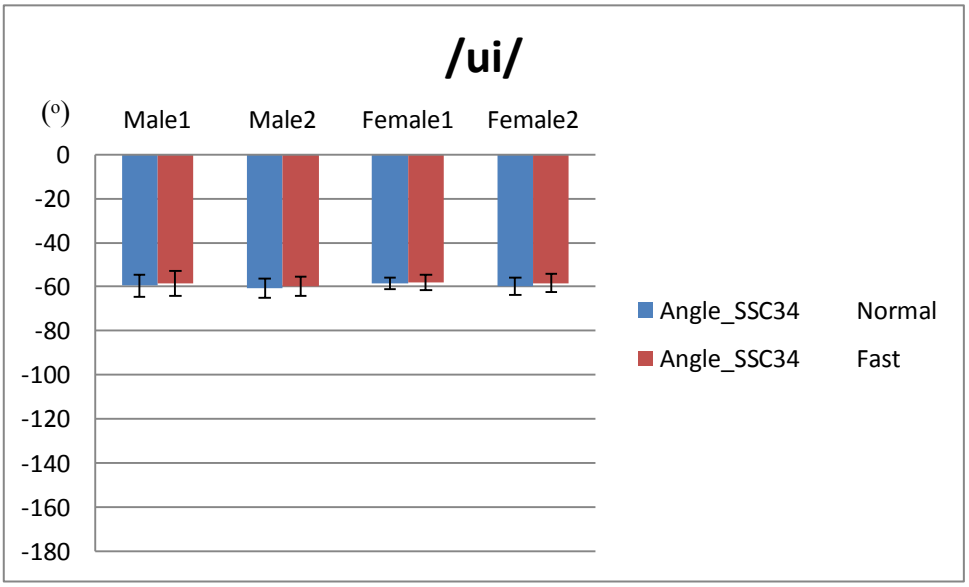


Figure 4-54: The average value and standard deviation of SSCF Angle23 of /ui/ produced by 4 French subjects (2 males + 2 females) at both normal and fast rate.

4.5.2.2.4 Other French V1V2 transition sequences

Following up the three SSCF Angles of the three French transition sequences /ai, ua, ui/ which are three boundaries of acoustic vowel triangle, the two same results are obtained. Firstly, for each transition sequence, the averages of each SSCF Angle are more or less the same with their very small standard deviation at both normal and fast rate. Secondly, the values for each SSCF Angle for each transition produced by all four French speakers are more or less the same. It means that the values of each SSCF Angle for transition sequences are independent with different speakers and are invariant among male and female speakers.

These characteristics are also observed for other transition sequences /aɛ, ae/ presented in Appendix 4.

4.5.2.3 French V1V2 transitions in 3D-plane of SSCF Angles

The SSCF Angles comparisons among five French transitions for each speaker were studied in section 4.5.2.1. SSCF Angles comparisons with the same transitions produced by four French native speakers (2 males + 2 females) also presented in section 4.5.2.2. In this section, we compare the different V1V2 transitions produced by the same four French native speakers by the combination of three SSCF Angles (SSCF Angle12, SSCF Angle23 and SSCF Angle34) in 3-D plane.

Figure 4-55 shows the French transitions of /ai, aɛ, ae, ua, ui/ produced by the same five French native speakers in 3-D plane which the x-, y-, and z-axis are SSCF Angle23, SSCF Angle12 and SSCF Angle34, respectively.

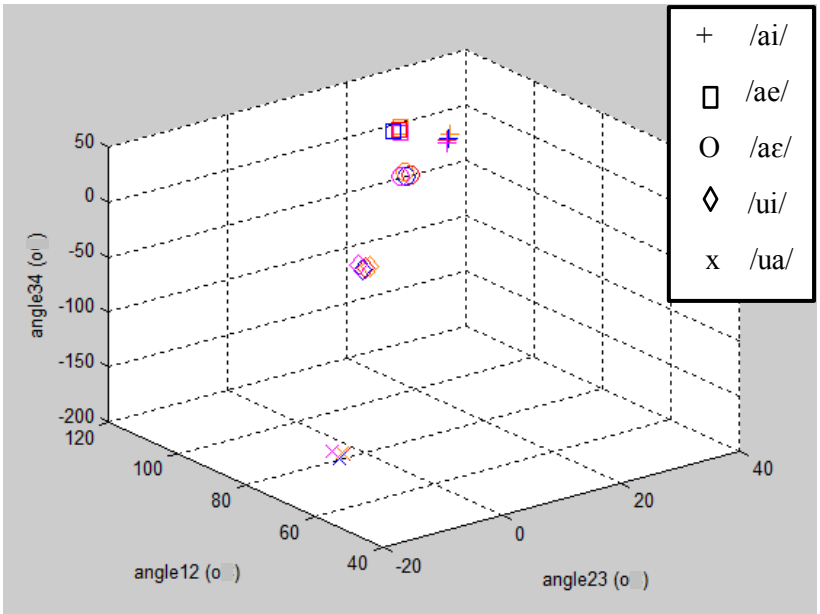


Figure 4-55: /ai, aɛ, ae, ua, ui/ transitions produced 4 French native speakers (2 males + 2 females) in 3-D plane of (SSCF Angle23, SSCF Angle12 and SSCF Angle34).

In this figure, the transitions of each speaker were displayed the same color, but with different markers, such as plus points were defined for /ai/ transitions; circle points were assigned for /aɛ/ transitions; and square points were defined for /ae/ transitions; diamond points were assigned for /ui/ transitions; and cross points were defined for /ua/ transitions.

Observing Figure 4-55, we found that firstly, the /ai/ transitions produced by four French native speakers (2 males + 2 females) converge in one region in this 3-D space. This corresponding result is also obtained with the /aɛ/, /ae/, /ua/, or /ui/ transitions. Secondly, five converging regions correspond to these five French transitions are completely separated mutually in the 3-D plane of SSCF Angles.

4.5.3 Conclusions

Five French transitions of /ai, aɛ, ae, ua, ui/ were compared in two cases: (i) first case is the SSCF Angles comparisons among different items for each speaker presented in section 4.5.2.1; (ii) second case is the SSCF Angles comparisons with the same item produced by four French native speakers (2 males + 2 females) presented in section 4.5.2.2. And finally, all five French transitions were presented in 3-D plane of SSCF Angles in section 4.5.2.3. Some interesting results were obtained: (i) for each speaker, each SSCF Angle is more or less invariant at both normal and fast speech rate for each French V1V2 transition; (ii) for each speaker, all five French transitions can be completely separated mutually in 3-D plane of SSCF Angles; (iii) for the same French V1V2 transitions, each SSCF Angle12, SSCF Angle23 or SSCF Angle34 is more or less the same value for four studied French native speakers which included 2 males and 2 females, in other words, each SSCF Angle is independent with different speakers; (iv) the transitions of the same French V1V2 transitions produced by four French speakers converged in one regions in 3-D space of SSCF Angles, and therefore, there were five distinct regions correspond to five studied transitions based on the combination of three SSCF Angle23, SSCF Angle12 and SSCF Angle34 in 3-D plane, in other words, the five French transitions of /ai, aɛ, ae, ua, ui/ produced by four French speakers were completed separated mutually.

Keeping up the analysis results in Vietnamese in section 4.4, we found that all these results in French transitions are similar to the analysis results of /ai, aɛ, ae, ua, ui/ transition in Vietnamese.

4.6 SSCF Angles comparisons between Vietnamese and French

In this section, we would like to compare the values of SSCF Angles of some V1V2 transitions in between Vietnamese and French. Is there any similarity in both languages? Is there any difference in both languages? We will consider this point in the following sections.

4.6.1 Stimuli

In this task, we will consider the five transitions (/ai, aɛ, ae, ua, ui/) in both Vietnamese and French. All these Vietnamese V1V2 transitions produced by eight Vietnamese native speakers (4 males +

4 females) that were described in detail in section 4.4.1.1. All five French V1V2 transitions produced by four French native speakers (2 males + 2 females) that were given in section 4.5.1.1. Each V1V2 sequence for both languages was recorded 5 times at normal rate and 5 times at fast rate.

4.6.2 Results

The following results show the comparisons of the same V1V2 transition in between French and Vietnamese.

4.6.2.1 /ai/ sequence

4.6.2.1.1 SSCF Angle12 of /ai/

Figure 4-56 presents the average and standard deviation of SSCF Angle12 of /ai/ transition produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate. It is interesting to realize that the averages of SSCF Angle12 of /ai/ transitions are positive and more or less the same values with the small standard deviations for both 8 Vietnamese speakers and 4 French speakers at both normal and fast speech rate.

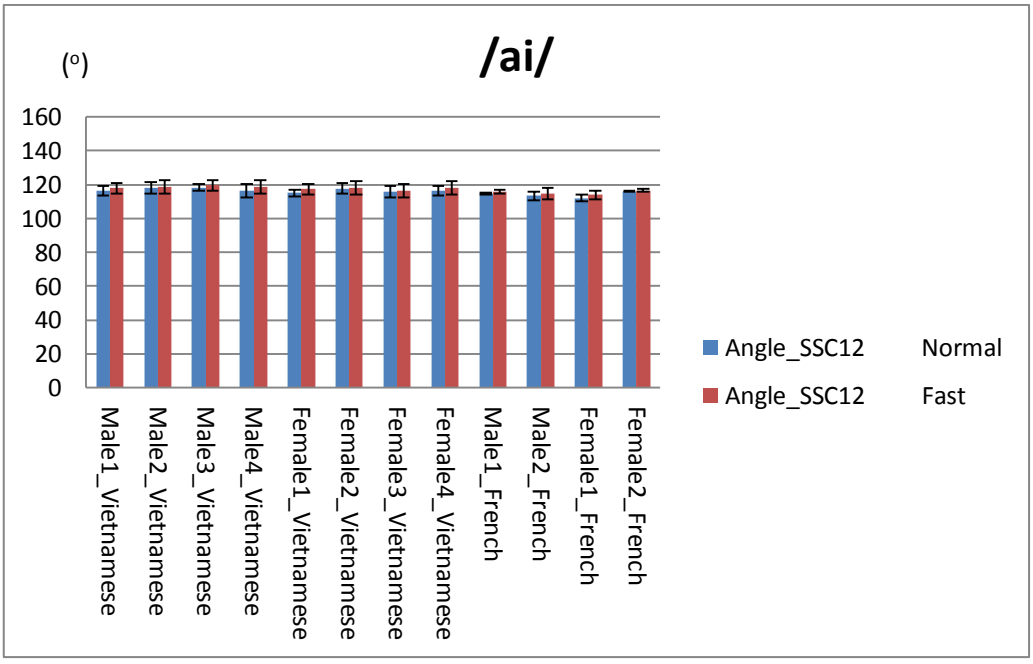


Figure 4-56: The average value and standard deviation of SSCF Angle12 of /ai/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rate.

4.6.2.1.2 SSCF Angle23 of /ai/

Figure 4-57 presents the average and standard deviation of SSCF Angle23 of /ai/ transition produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native

males + 2 French native females) at both normal and fast rate. All the averages of SSCF Angle12 of /ai/ transitions are positive, but the SSCF Angle12 values of /ai/ transitions for French are a little higher than the one for Vietnamese.

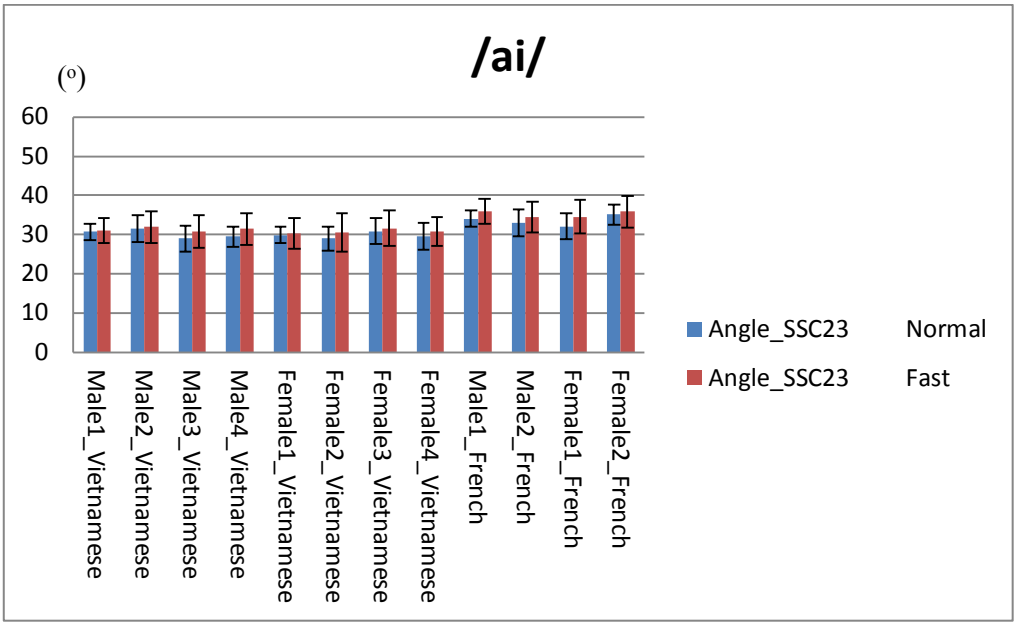


Figure 4-57: The average value and standard deviation of SSCF Angle23 of /ai/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rate.

4.6.2.1.3 SSCF Angle34 of /ai/

The averages and standard deviations of SSCF Angle34 of /ai/ transition produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate were presented in Figure 4-58.

It is clear to realize that there is a difference among the SSCF Angle34 values of /ai/ transitions in two languages, namely, the absolute values of SSCF Angle34 of /ai/ transitions in French are much higher than the one in Vietnamese.

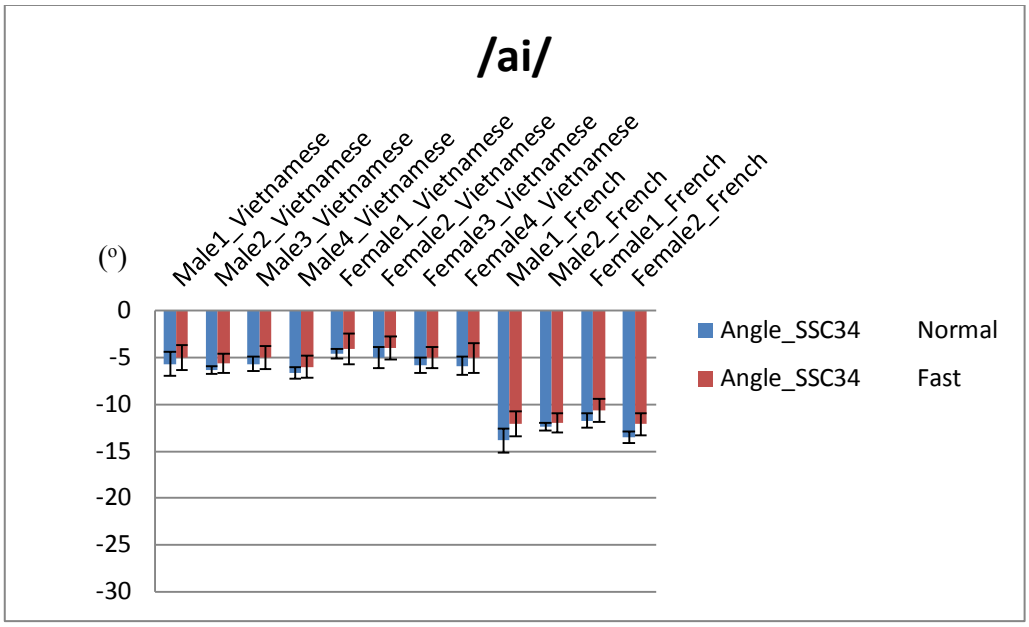


Figure 4-58: The average value and standard deviation of SSCF Angle34 of /ai/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rate.

4.6.2.2 /aɛ/ sequence

4.6.2.2.1 SSCF Angle12 of /aɛ/

Figure 4-59 presents the average and standard deviation of SSCF Angle12 of /aɛ/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate. We found out that the averages of SSCF Angle12 of /aɛ/ transitions are positive and more or less the same values with the small standard deviations for both 8 Vietnamese speakers and 4 French speakers at both normal and fast speech rate.

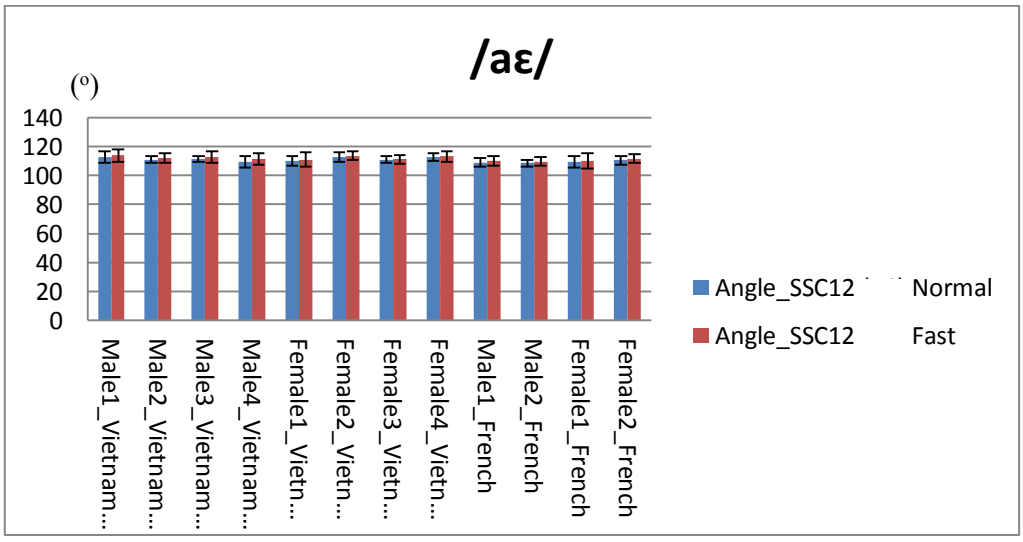


Figure 4-59: The average value and standard deviation of SSCF Angle12 of /aɛ/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rate.

4.6.2.2.2 SSCF Angle23 of /aɛ/

The averages and standard deviations of SSCF Angle23 of /aɛ/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate were presented in Figure 4-60.

There is a little difference among the SSCF Angle23 values of /aɛ/ transitions in two languages, namely, the absolute values of SSCF Angle23 of /aɛ/ transitions in French are smaller than the one in Vietnamese.

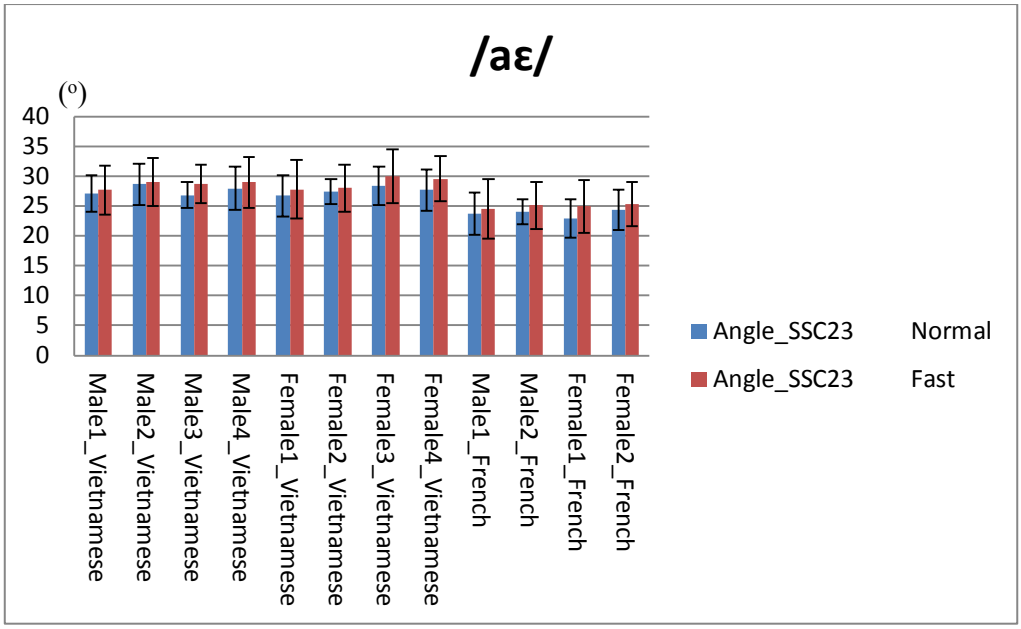


Figure 4-60: The average value and standard deviation of SSCF Angle23 of /aɛ/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rate.

4.6.2.2.3 SSCF Angle34 of /aɛ/

Figure 4-61 presents the averages and standard deviation of SSCF Angle34 of /aɛ/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate.

As a whole, all the averages of SSCF Angle34 values of /aɛ/ transitions are negative, but the absolute values of SSCF Angle34 of /aɛ/ transitions for French are a little lower than the one for Vietnamese.

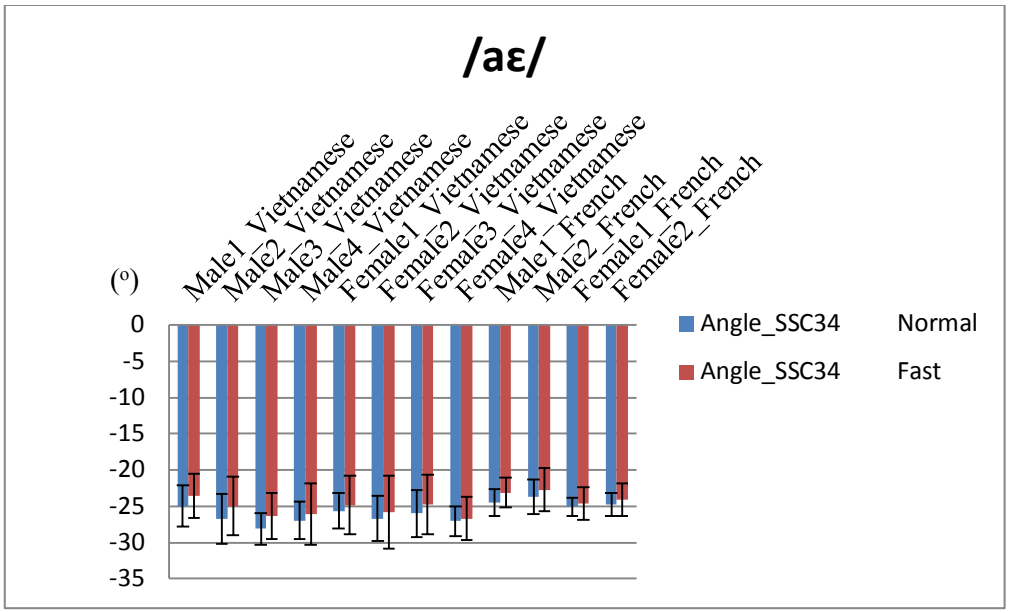


Figure 4-61: The average value and standard deviation of SSCF Angle34 of /aɛ/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

4.6.2.3 /ae/ sequence

4.6.2.3.1 SSCF Angle12 of /ae/

Figure 4-62 presents the averages and standard deviations of SSCF Angle12 of /ae/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate.

We observed that the averages of SSCF Angle12 of /ae/ transitions are positive and more or less the same values with the small standard deviations for both 8 Vietnamese speakers and 4 French speakers at both normal and fast speech rate.

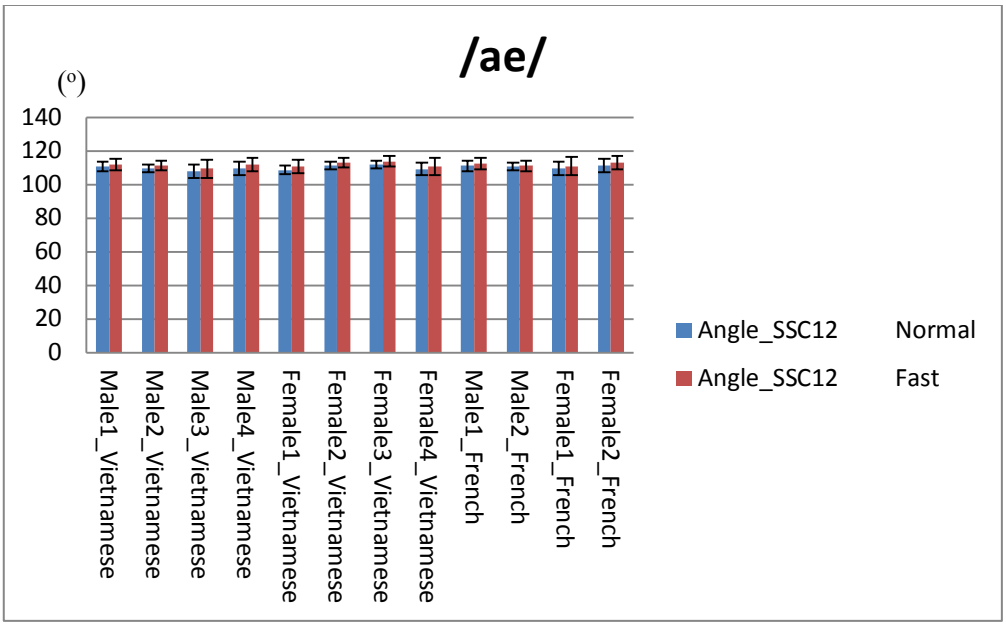


Figure 4-62 The average value and standard deviation of SSCF Angle12 of /ae/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

4.6.2.3.2 SSCF Angle23 of /ae/

Figure 4-63 presents the averages and standard deviations of SSCF Angle23 of /ae/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate.

We observed that the averages of SSCF Angle23 of /ae/ transitions in French for both males and females are higher than the one in Vietnamese.

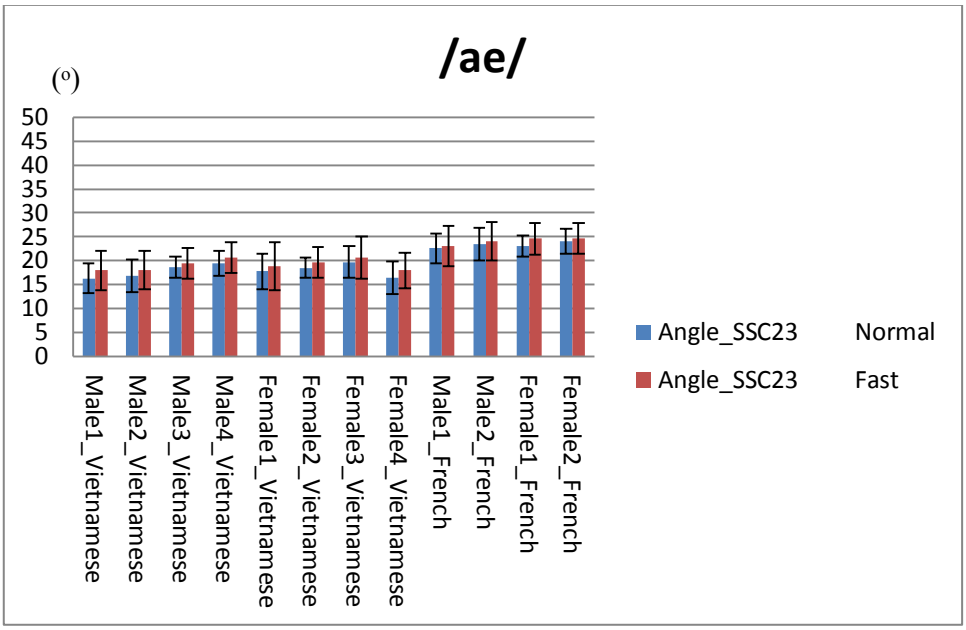


Figure 4-63: The average value and standard deviation of SSCF Angle23 of /ae/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

4.6.2.3.3 SSCF Angle34 of /ae/

The averages and standard deviations of SSCF Angle34 of /ae/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate were presented in Figure 4-64.

In this case, the values of SSCF Angle34 of /ae/ transitions in French with both males and females are higher than the one in Vietnamese.

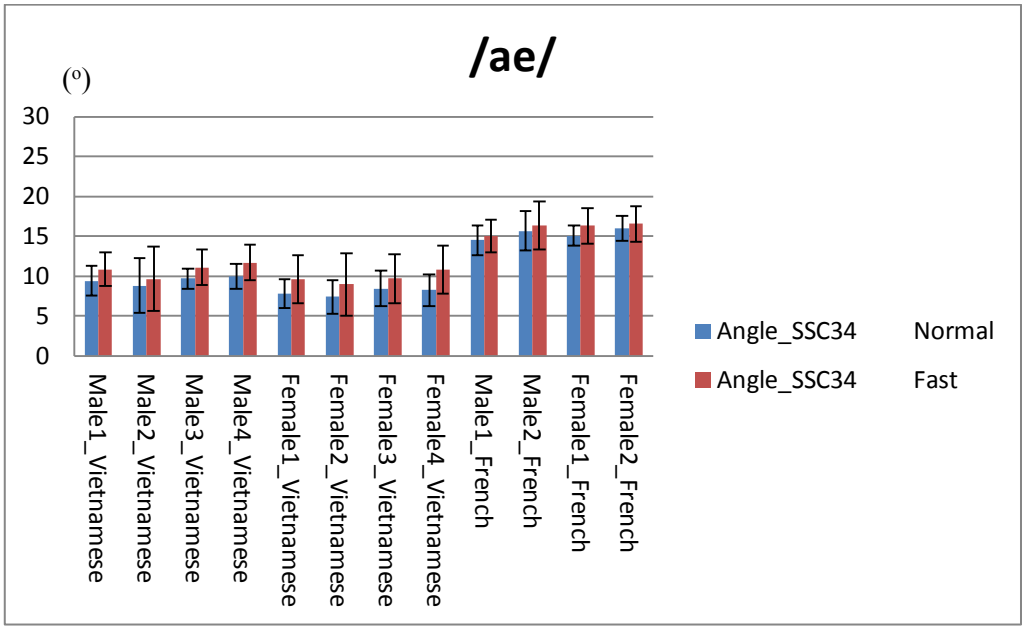


Figure 4-64: The average value and standard deviation of SSCF Angle34 of /ae/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

4.6.2.4 /ua/ sequence

4.6.2.4.1 SSCF Angle12 of /ua/

Figure 4-65 presents the averages and standard deviations of SSCF Angle12 of /ua/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate.

We found out that the averages of SSCF Angle12 of /ua/ transitions are positive and more or less the same values with the small standard deviations for both 8 Vietnamese speakers and 4 French speakers at both normal and fast speech rate.

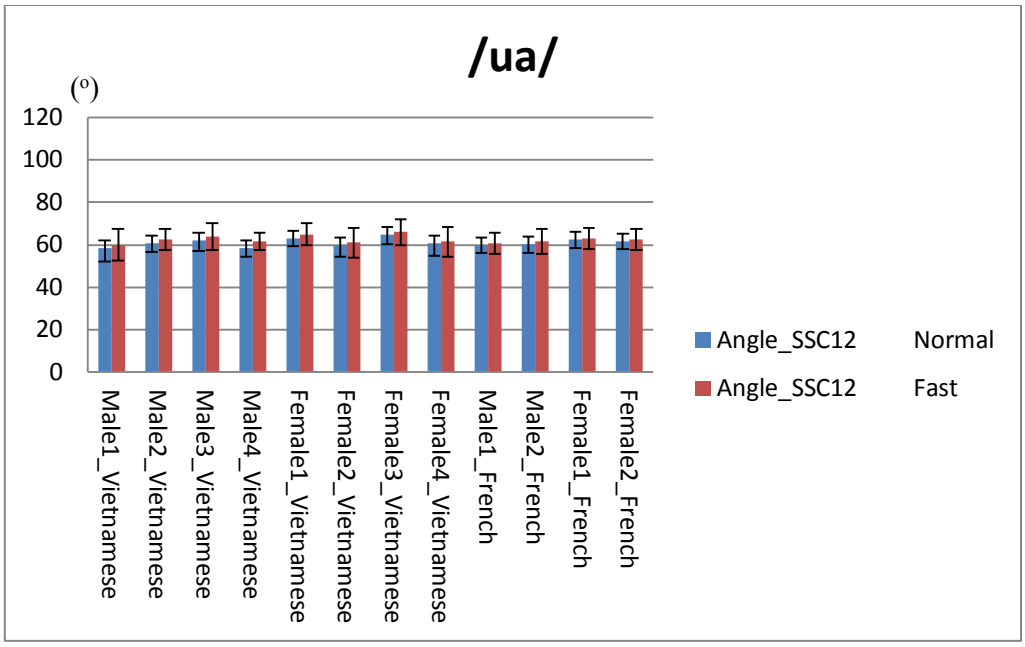


Figure 4-65: The average value and standard deviation of SSCF Angle12 of /ua/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

4.6.2.4.2 SSCF Angle23 of /ua/

Figure 4-66 presents the averages and standard deviations of SSCF Angle23 of /ua/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate.

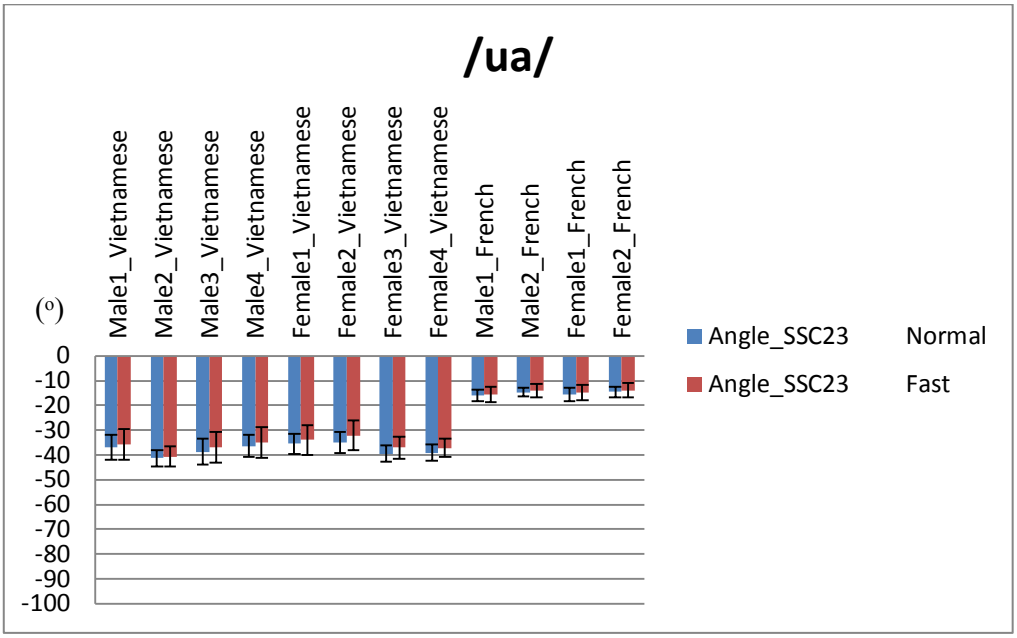


Figure 4-66: The average value and standard deviation of SSCF Angle23 of /ua/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

We realized that the averages of SSCF Angle23 of /ua/ transitions are negative, but the absolute vales of SSCF Angle23 of /ua/ transition in French for both males and are much smaller than the one in Vietnamese.

4.6.2.4.3 SSCF Angle34 of /ua/

The averages and standard deviations of SSCF Angle34 of /ua/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate were presented in Figure 4-67.

We observed that there is a little difference among the averages of SSCF Angle23 of /ua/ transitions in French for both males and the one in Vietnamese.

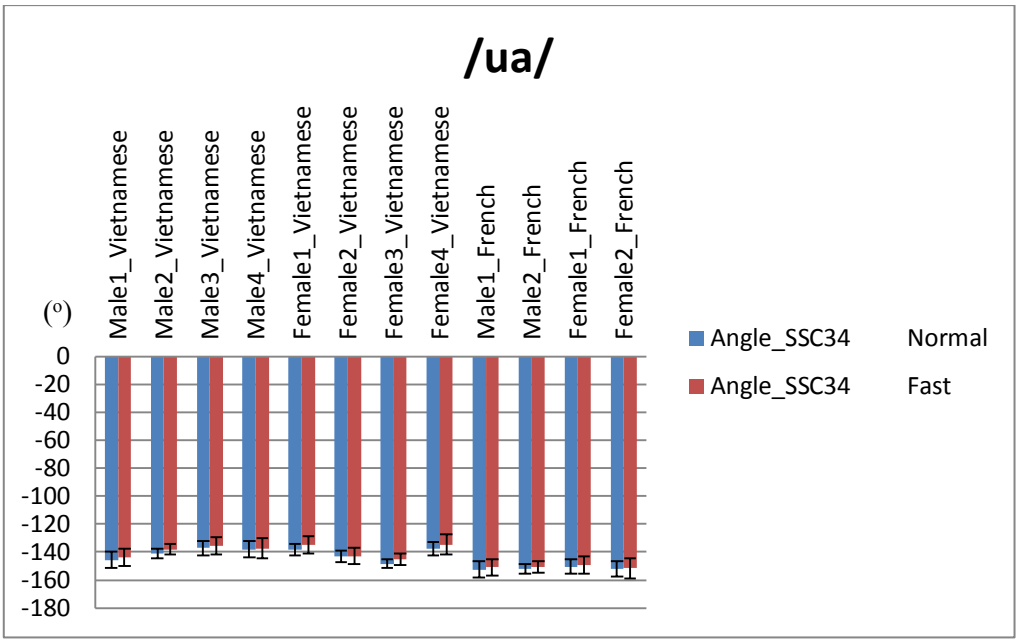


Figure 4-67: The average value and standard deviation of SSCF Angle34 of /ua/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

4.6.2.5 /ui/ sequence

4.6.2.5.1 SSCF Angle12 of /ui/

Figure 4-68 presents the averages and standard deviations of SSCF Angle12 of /ui/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate.

We found out that the averages of SSCF Angle12 of /ui/ transitions are positive and more or less the same values with the small standard deviations for both 8 Vietnamese speakers and 4 French speakers at both normal and fast speech rate.

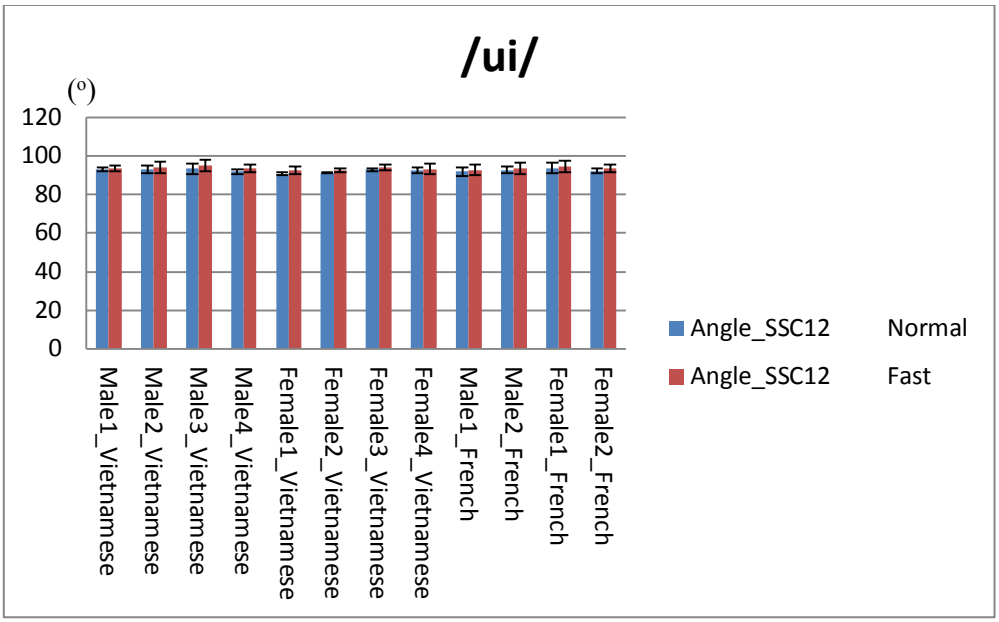


Figure 4-68: The average value and standard deviation of SSCF Angle12 of /ui/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

4.6.2.5.2 SSCF Angle23 of /ui/

Figure 4-69 presents the averages and standard deviations of SSCF Angle23 of /ui/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate. We realized that the averages of SSCF Angle23 of /ui/ transitions in French for both males and females are much higher than the one in Vietnamese.

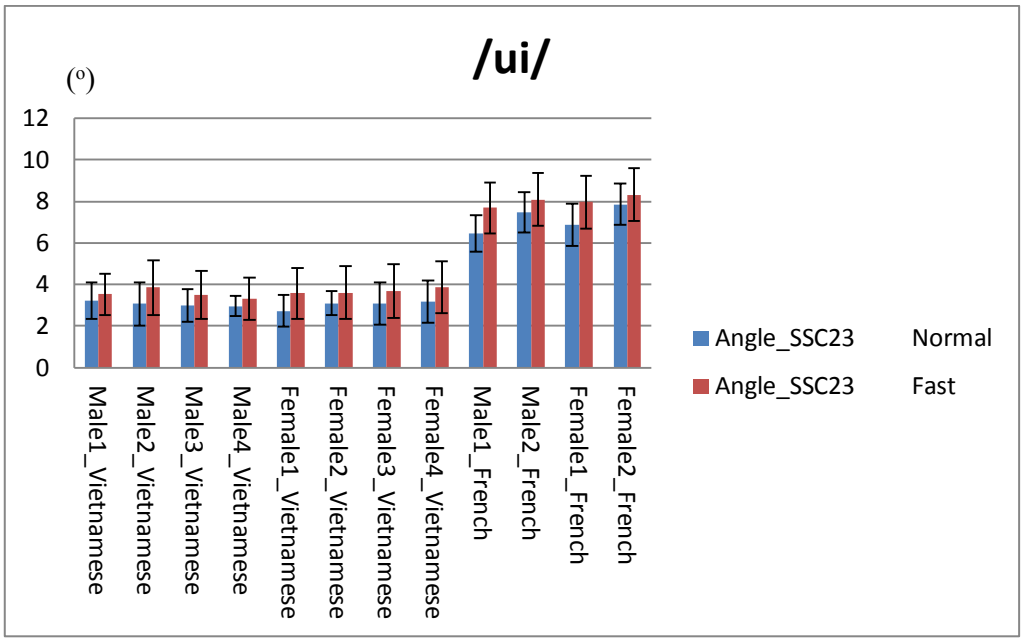


Figure 4-69: The average value and standard deviation of SSCF Angle23 of /ui/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

4.6.2.5.3 SSCF Angle34 of /ui/

The averages and standard deviations of SSCF Angle34 of /ui/ transitions produced by twelve speakers (4 Vietnamese native males + 4 Vietnamese native females + 2 French native males + 2 French native females) at both normal and fast rate were presented in Figure 4-70. We observed that the averages of SSCF Angle34 of /ui/ transitions in French for both males and females are much smaller than the one in Vietnamese.

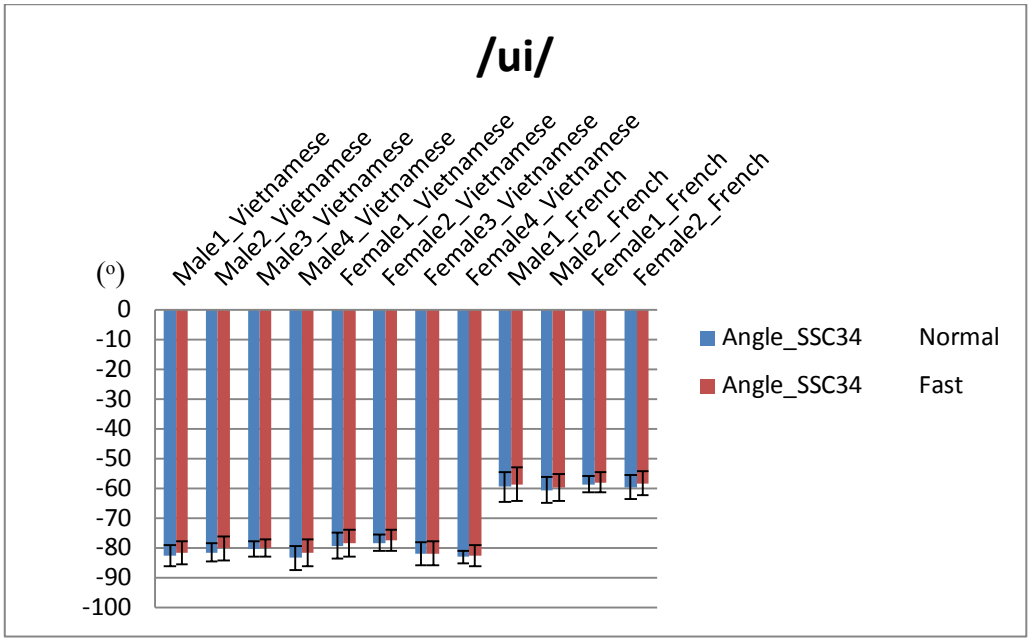


Figure 4-70: The average value and standard deviation of SSCF Angle34 of /ui/ produced by 12 speakers (4 Vietnamese males + 4 Vietnamese females + 2 French males + 2 French females) at both normal and fast rates.

4.6.3 Discussions

Keeping up the obtained results in section 4.6.2, we found that there are some similarities and also some differences in the comparisons of SSCF Angles of the same V1V2 transition in between Vietnamese and French.

4.6.3.1 The similarities on SSCF Angles between Vietnamese and French

Following up the obtained results from section 4.4.2, 4.5.2 and 4.6.2, we found out the similar points for each language Vietnamese or French: (i) for each speaker, each SSCF Angle (SSCF Angle12, SSCF Angle23 and SSCF Angle34) of the same V1V2 transition is more or less the same in both normal and fast rate; (ii) each SSCF Angle of the same V1V2 transition is more or less the same for both male and female speakers; (iii) the combination of three SSCF Angles (SSCF Angle12, SSCF Angle23 and SSCF Angle34) brought out the useful characteristic in order to completely separate the different V1V2 transitions.

4.6.3.2 The differences on SSCF Angles between Vietnamese and French

Comparing SSCF Angles in between Vietnamese and French in five transitions of /ai, aɛ, ae, ua, ui/, we found that: (i) the values of SSCF Angle12 for the same V1V2 transition are more or less the same in two languages Vietnamese and French; (ii) but there are the differences on the values of the SSCF Angle23 or SSCF Angle34 in the same V1V2 transition in between Vietnamese and French. This point showed that although two languages Vietnamese and French have some similarities in vowel system, such as /a, ɛ, e, i, u/, but there are the differences on acoustic characteristics on the same V1V2 transition. These differences need to study further in order to obtain the more detail explanations.

4.7 Conclusions of chapter 4

Chapter 4 studied the modeling of dynamic acoustic speech features. First step was presented in section 4.1, SSCF parameters were proposed as "pseudo-formant" parameters. Some comparisons with formant frequencies were performed, and finally, SSCF parameters were considered to be similar to formant frequencies but contrary to formant frequencies, they are continuous even during the obstruent consonant production.

Next, based on SSCF parameters, in section 4.2, we proposed a new way to model the dynamic acoustic speech features and we called SSCF Angles. The method to compute SSCF Angles was presented in detail in section 4.3.

After that, some analysis work on SSCF Angle parameters were performed on some transitions of Vowel-to-Vowel (V1V2) of both Vietnamese and French in section 4.4 to 4.6. SSCF Angles are precise and reliable parameters. The analysis results showed that the V1V2 transitions can be completely separated based on SSCF Angles in both Vietnamese and French. V1V2 and V2V1 have symmetrical properties on acoustic domain based on SSCF Angles.

The analysis results on SSCF Angles also showed that each SSCF Angle is more or less the same value for both male and female speakers on the same V1V2 transition sequence. And they are also fairly invariant for speech rate (normal and fast speech rate) for each speaker.

In this basis, we propose to verify the usable possibility and independent properties with speakers (for both males and females) of SSCF Angles by using automatic Vietnamese speech recognition system. This task will be considered in Chapter 5.

Chapter 5
Using dynamic acoustic speech features in automatic speech recognition

5.1 Introduction

An encouraging result from Chapter 4 is that the angle parameters proposed in this dissertation (SSCF angles) are fairly speaker-independent. They thus appear as potentially promising candidates for integration into speech recognition systems.

The expected gains relate crucially to the fact that a speech recognition system built on this basis would be much less speaker-dependent: a smaller number of speakers would be required to arrive at a sufficient and adequate sample. Hence, this would decrease dramatically the amount of data that is necessary for training the ASR system.

Keeping in mind that our study is still exploratory in nature – we have studied the dynamics of acoustic gestures only for vowel-vowel transitions, and consonant-vowel transitions will need to be addressed in future –, we nonetheless wish to attempt a set of tests concerning the use of angle parameters in speech recognition.

For this purpose, we decided to use a simple and very classic automatic speech recognition system for Vietnamese language that was developed at the MICA Institute.

The goal is not to arrive at an excellent recognition system as yet, but to see how our proposed SSCF Angle parameters compare with the MFCC parameters conventionally used in speech recognition.

We decided to divide our tests into two parts: in the first set of tests, the training data consisted of male and female voices pooled together, as did the testing data, i.e. the classic setup for the less speaker-dependent ASR. In the second, a model trained on female voices was applied to the recognition of male voices, and vice versa – a more challenging task in terms of achieving speaker-independent (gender-independent) ASR.

As a preliminary to an account of these experiments, some reminders about the fundamentals of speech recognition are offered in the following paragraphs. Readers with a speech processing background can safely proceed straight to section 5.2.3.

5.2 Description of the used ASR system

Scientific studies on speech recognition began in early 1950, but the first tangible results was gave in the 70s with the project ARPA (Advanced Research Projects Agency) (Newell, 1973; Klatt, 1977). Their main goals was to design a robust and efficient speech recognition system for integrating naturally in communication between human and machine.

Speech recognition applications are becoming more and more useful nowadays. Example, in disability community, voice recognition can help people with musculoskeletal disabilities caused by multiple sclerosis, or arthritis achieve maximum productivity on computers.

5.2.1 Classification of ASR system

Speech recognition systems can be separated in several different classes depending on what types of utterances they have the ability to recognize, or on what kinds of speaker model, or on what types of vocabulary size.

5.2.1.1 Type of ASR system based on utterances

Some following types of ASR are classified depending on utterances:

Firstly, isolated words: Isolated word recognition system recognizes single word. It is suitable for situations where the user is required to give only one word response or commands, but it is very unnatural for multiple word inputs.

Secondly, connected words: A connected words system recognizes sequences of some words without voluntary pausing among them, example, recognition of connected digits or numbers, etc.

Thirdly, continuous speech: Continuous speech recognition system allows users to speak almost naturally, while computer determines its content. Basically, it is computer dictation which the closest words run together without pause or any other division between words.

Finally, spontaneous speech: Spontaneous speech recognition system recognizes the natural speech that comes suddenly through mouth.

5.2.1.2 Type of ASR system based on speaker model

Based on speaker model, ASR system can be classified into two following systems:

The first one is speaker dependent model. Speaker dependent system are developed for a particular type of speaker. They are more accurate for the particular speakers, but could be less accurate for other speakers.

The second one is speaker independent model. Speaker independent system can recognize a variety of speakers without any prior training. This system is developed to operate for any particular type of speaker.

The third one is speaker adaptive model. Speaker adaptive speech recognition system uses the speaker dependent data and adapt to the best suited speaker to recognize the speech and decreases error rate by adaptation.

5.2.1.3 Type of ASR based on vocabulary

Depending on the size of vocabulary, three ASR system were defined as following:

- ASR with small vocabulary includes 1 to 100 words;
- ASR with medium vocabulary contains 101 to 1000 words;
- ASR with large vocabulary consists of more than 1000 words.

5.2.2 Brief review of the structure of an ASR system

Automatic speech recognition (ASR) means an automated process that inputs human speech and tries to find out what was said. In other words, an ASR system converts the speech signal into words. The recognized words can be (i) a final output (an end product, as it were), or (ii) the input to further natural language processing. Figure 5-1 illustrates the major modules of an ASR system (Rabiner and Juang, 1993; Glass, 2007).

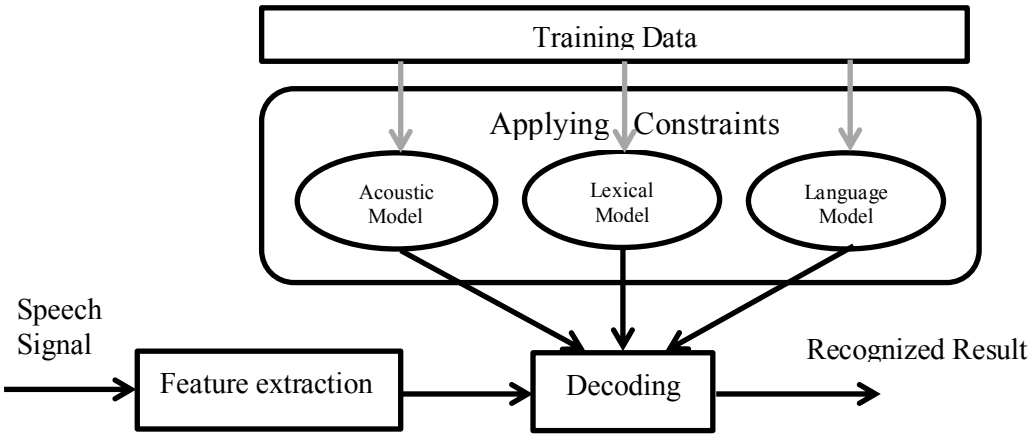


Figure 5-1: Major components in an automatic speech recognition system (Rabiner and Juang, 1993;Glass, 2007).

In feature extraction, signal processing techniques are applied to the speech signal in order to dig out the features that distinguish different phonetic unit from each other. A phonetic unit can be a

phoneme, a syllable or a word. Given the features extracted from the speech, acoustic model provides probabilities for different phonetic units at different time instants.

Acoustic model is the connection between acoustic information and phonetics. Training process establishes co-relation between the basic speech units and the acoustics observations. Training of the ASR system requires creating a pattern representative for the features of class using one or more patterns that correspond to speech sounds of the same class (Saksamudre et al., 2015).

Lexical model is used to create words from the acoustic model. Language model contains the structural constraints available in the language to generate the probabilities of occurrence. It includes the probability of a word occurrence after a word sequence. Each language has its own constraints. The language model distinguishes words and phrases that have similar sound.

The acoustic model, lexical model and language model are used in decoding for searching the recognition hypothesis that fits best to the models. Recognition result can then be used in various applications.

Selection of proper acoustic features is one of the important tasks in the design of every system using speech processing in common, and in ASR system in particularly.

However the acoustic characteristics of speech still suffer from some limitations: indeed, for some 40 years, speech is considered as a sequence of quasi-stable signals (vowels) separated by transitions (consonants). It is also commonly accepted that vowels are acoustic "targets" that must be reached to make speech understandable. But in fact, vowels in fluent speech are rarely steady-state, even well-known standard formant frequencies targets used for vowel description cannot be regarded as invariant since they exhibit significant variations across speakers (male, female, adult/child subjects), and across languages.

In the latter half of the twentieth century, some researchers recognized that natural speech is not a simple sequence of steady-state segments. In other words, speech is a dynamic process. Our studies presented in the previous chapters also contributed to show the role of acoustic dynamics of speech features. Thus, in Chapter 5, we would like to apply some dynamic speech features on Vietnamese ASR and to evaluate the ASR performance.

5.2.3 Classical approach using MFCC parameters

MFCC (Mel Frequency Cepstral Coefficient) is the most popular speech feature currently used on automatic speech recognition system in clean environment with high performance. They were introduced by Davis and Mermelstein (1980) (from which the information summarized in this paragraph is taken), and have become state-of-the-art ever since.

MFCC parameters will be calculated as following algorithm in Figure 5-2. The first stage in MFCC extraction is pre-emphasis to boost the amount of energy in the high frequencies of speech signal. Then the stationary portion of speech will be extracted by using a window. A more common window used in MFCC extraction is the Hamming.

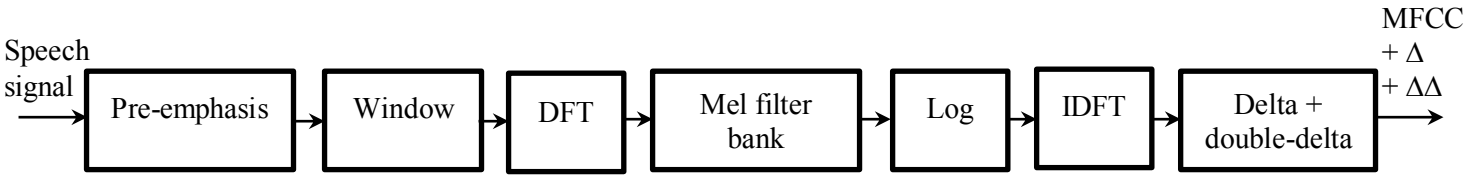


Figure 5-2: MFCC extraction algorithm (Davis and Mermelstein, 1980).

The next step is to extract spectral information for these windowed signals by discrete Fourier transform (DFT). The output result of the DFT will bring the information about the amount of energy at each frequency band, and after that, will be passed a non-liner filter in Mel scale. And the Figure 5-3 illustrates the general form of this filter bank. As can be seen, the filters are used triangular and the mel frequency *m* can be defined as following:

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad \textbf{(5-1)}$$

where f is the frequency in Hz, m is mel frequency of f.

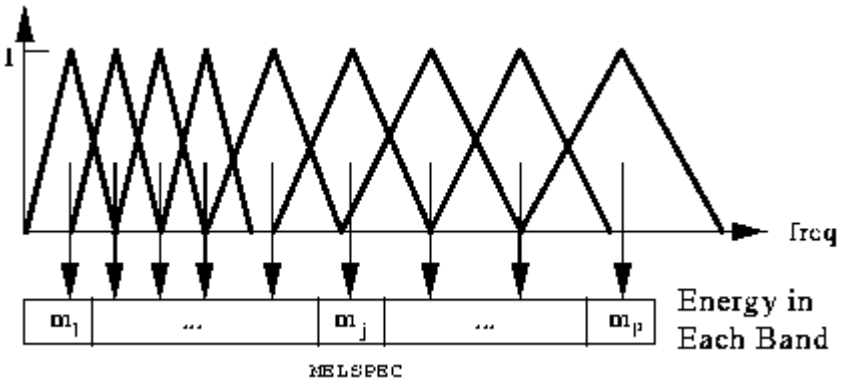


Figure 5-3: Mel-Scale filter bank.

Then, we take the log of the mel spectrum values. The next step is the computation of the cepstrum. These cepstrum coefficients are more formally defined as the inverse DFT of the log magnitude of the DFT of a signal. These cesptrum coefficients are called MFCC coefficient.

Therefore, MFCCs are a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency (that approximates the human auditory system's response), and finally converted to cepstral representation. MFCCs are static parameters. In current ASR, 13 MFCC parameters are considered as the best number of speech features for high ASR performance until now.

In fact, because speech signal is not constant from frame to frame, we also add features related to the change in cepstral features over time by computing their delta (Δ) and double-delta coefficients (ΔΔ).

Delta and delta-delta of MFCCs are the speed and acceleration parameters derived from MFCCs. They were considered as the first dynamic parameters. Therefore, there are 13 delta and 13 delta-delta parameters corresponding to 13 MFCCs.

The delta coefficients are calculated as formula 5-2.

$$\Delta_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad \textbf{(5-2)}$$

where Δ_t is a delta coefficient, from frame t computed in terms of the static MFCC coefficients c_{t+N} to c_{t-N} . A typical value for N is 2. Double delta are calculated in the same way, but they are calculated directly from the deltas.

Disadvantage of MFCCs: although MFCC coefficients showed good performance in speech recognition system, but they are still sensitive to channel mismatches between training and testing data, and they are also speaker dependent (El-Samie, 2011).

5.2.4 Decoding

Decoding is the process of comparing the unknown test pattern with each sound class reference pattern and computing a measure of similarity between them. There exist some following approaches to speech recognition.

5.2.4.1 Template-based approach

Template based approach has a collection of prototypical speech patterns (Rabiner and Juang, 1993). These patterns are stored as reference patterns representing the dictionary of words. Speech is recognized by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern. Normally, templates for entire words are constructed. This approach is simple. But this method has the drawback that the pre-recoded templates are fixed, so variations in speech signals can only be modeled by using many templates per word. This is certainly impractical and continous speech recognition is not possible using this method.

5.2.4.2 Stochastic approach

Stochastic modeling entails the use of probabilistic models to deal with uncertain or incomplete information, example in speech recognition, from variability's speaker, contextual effects, etc. Stochastic models are particularly suitable approach to speech recognition. The most popular

stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state markov model and a set of output distributions. The transition parameters in the Markov chain models, temporal variabilities, while the parameters in the output distribution model, spectral variabilities.

Therefore, the parameters of the model are the state transition probabilities, the means, variances and mixture weights that represent the state output distributions (Rabiner and Juang, 1993). Each word or phoneme will have a different output distribution. HMM can used for a large vocabulary speech recognition system. But the main drawback of statistical models is that they must make a priori modeling presumption, which is liable to be inaccurate, restrict the system's performance.

5.2.4.3 Dynamic Time Warping (DTW)

Dynamic Time Warping is an algorithm for measuring similarity between two sequences which may vary in time or speed (Rabiner and Juang, 1993). In general, DTW is a method that allows a computer to find an optimal match between two given sequences. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. One example of the restrictions imposed on the matching of the sequences is on the monotonicity of the mapping in the time dimension.

5.2.4.4 Knowlege-based approach

This method uses the information regarding linguistic, phonetic and spectrogram (Saksamudre et al., 2015).. The expert knowledge about variation in speech is hand-coded into a system. It takes set of features from the speech and then train the system to generate set of production rules automatically from the samples. These rules are resulted from the parameters that provide useful information about a classification. The effort of recognition is performed at the frame level, using an inference engine to implement the decision tree and classify the firing of the rules.

This approach has the benefit of explicitly modeling variation in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully, so this approach is considered as impractical.

5.2.4.5 Neural network based approach

Neural network takes several frames of coefficients as input and produces posterior probabilities over HMM states as output. This method can obtain more accuracy than HMM based systems if the training data and vocabulary size is limited (Saksamudre et al., 2015). Normally, neural networks are used for phoneme recognition. And there are also an NN-HMM hybrid system that combine between neural network and HMM, in which, the neural network as part of phoneme recognition and the HMM as part of language modeling.

Deep neural networks (DNNs) means using a neural network with many layers of non-linear hidden units and a very large output layer. The series of layers between input and output do feature identification and processing in a series of stages, just as our brains seem to. This new machine learning algorithms can lead to significant advances in automatic speech recognition (Hinton et al., 2012).

5.2.5 Vietnamese speech corpus

For this task, we used an available Vietnamese read-speech corpus: VNSpeechCorpus (Le et al., 2004). The texts were tailored to the needs of ASR systems; they contain five kinds of data: (i) phoneme; (ii) tones; (iii) digits and string of digits; (iv) application words; (v) sentences and paragraphs.

The speakers were employees of the International Research Institute MICA, teachers and students of Hanoi University of Science and Technology and their friends. There are 32 Vietnamese speakers (16 males and 16 females) in the current VNSpeechCorpus version stored at MICA Institute. They were from major cities and provinces, such as Hanoi, Nghe An, Ha Tinh, Ho Chi Minh city, which represented three major dialect regions: the South, the North, and the Centre, respectively. Their ages range from 22 to 40 years old.

VNSpeechCorpus was recorded by EMACOP system developed at CLIPS-IMAG Laboratory (France). EMACOP was a multimedia environment for acquiring and managing speech corpora, running under Windows 9x and Windows NT. All these recordings took place in the recording booth of MICA Institute, Hanoi, Vietnam, with a SENNHEISER HMD 410-6 head microphone and microphone pre-amplifiers Soundcraft Spirit Folio FX8. The sampling frequency was 16 kHz (16 bits). Each speaker were recorded about 60 minutes, including 45 common minutes of phonemes, tones, digits and strings of digits, application words and common sentences and paragraphs corpus, and 15 private minutes of about 40 short paragraphs.

5.2.6 The MICA ASR system

Automatic Vietnamese speech recognition system was deployed on the corpus in section 5.3.1.1, using Sphinx toolkit of Carnegie Mellon University (CMU Sphinx, 2015). Sphinx is a ASR system using statistical methods to train models. Hidden Markov Models (HMMs), decision trees are used throughout the system (Jelinek, 1976; Rabiner and Juang, 1993; Steve Young, 1996; Huang et al., 2001).

The ASR's quality here is evaluated by Syllable Error Rate (SER) factor in percent (%). The smaller SER the ASR system has, the better its quality is.

In our experiments conducted in order to test the dynamic acoustic speech features in ASR, including our proposed dynamic acoustic speech features – SSCF Angle parameters, we propose two types of ASR: (i) one using classical MFCC parameters; (ii) one using SSCF angle parameters as ASR's input. For each type of ASR, the performances of ASR using different single parameter sets were compared and finally, we performed comparisons between the two types. The following experiments 1 to 4 were carried out in succession.

5.3 Experiment 1: Role of dynamic acoustic speech features in automatic Vietnamese speech recognition system

Over the past twenty years, numerous works on speech recognition have shown that, whatever the spectral parameters used to characterize the speech signal, it is necessary to supplement them by their first and second derivatives, respectively named "delta" and "delta-delta (Rabiner and Juang, 1993; Huang et al., 2001).

These derivative parameters characterize the speed and the acceleration of the temporal evolution of each spectral parameter. So, this is a rough-and-ready way of characterizing the dynamics of the speech signal and this characterization is done only on the signal itself, without taking in consideration the proper dynamic of the speech production model.

In this section, we propose to evaluate the role of delta and/or delta-delta of classical MFCCs parameters in our Vietnamese speech recognition system. It is a way for us to better understand the importance of dynamic characteristics of the speech signal in automatic speech recognition (as in speech perception, which is no other than speech recognition by humans).

5.3.1 Methodology

The Vietnamese database set was selected from the current VNSpeechCorpus version at MICA Institute (in section 5.2.3). All 26 Vietnamese native speakers were included (remember that they cover dialectal diversity from the North, the South and the Middle of Vietnam). As for materials, they include both common and private paragraph. This corpus is divided into two sets as follows:

- Training set includes 11 hours of speech from 26 Vietnamese speakers which contains 13 males and 13 females.
- Testing set consists of 2.5 hours from 6 Vietnamese speakers which includes 3 males and 3 females.

For this initial test, our ASR system only uses acoustic model (without language model) and Vietnamese ASR non-tonal phonemes. For instance, in the sentence "Trời đẹp quá" [tɕɤj dɛ̂p kwa], the

following non-tonal phonemes were found: /tɕ/, /ɤ/, /j/, /đ/, /ɛ/, /p/, /k/, /w/ and /a/. Each phoneme is modeled by a HMM model.

5.3.2 Results and discussions

Table 5-1 presents the results of syllable error rate of Vietnamese ASR in this experiment using MFCCs and their derivations (Delta and/or Delta-Delta). We call ASR in this experiment 1, ASR1. In this figure, Δ, ΔΔ are defined respectively for Delta and Delta-Delta parameters of MFCCs.

Table 5-1: Syllable Error Rate (SER) of ASR1 using MFCCs and their derivations: Δ (Delta parameters of MFCCs); ΔΔ (Delta-Delta parameters of MFCCs).

Type of input parameters	SER (%)	Dimension of the parameter vector
MFCC + Δ + ΔΔ	25.302	39
Δ + ΔΔ	26.572	26
MFCC	35.869	13
Δ	32.338	13
ΔΔ	66.207	13

Reminding MFCCs are static speech features, and delta and/or delta-delta parameters are the dynamic parameters.

Observing Table 5-1, we find that the best result (characterized by the smallest SER score) is obtained by the classical parameter vector made of the three sets of MFCC, Delta and Delta-Delta parameters: SER = 25.302 %. This is good news in terms of the hypothesis tested – that dynamic parameters can be usefully applied in speech recognition.

In comparison with performances obtained by usual recognition systems proposed on the market, the score is not brilliant. However, we recall that our system does not use a language model and also that our training database (corpus) is reduced as it contains a small number of speakers. In addition, in this test we do not take tones into account. There was thus absolutely no hope that this test model would outperform fine-tuned commercial ASR software currently available for Vietnamese.

Results presented in Table 5-1 allow for two comparisons, as follows:

Firstly, comparing the three scores using only a single set of 13 parameters, we find that the score obtained with MFCC parameters only and that obtained with Δ parameters only are similar. For MFCCs, SER score is 35.869 % and for Δ, it is 32.338 %. However the score obtained by the ASR using only Δ parameters is better, with a lower SER of 3.531 %.

This result seems to show that the information given by the Δ parameters are more efficient than the corresponding initial parameter MFCC. Of course, we have to take precautions because the training database of our recognition system is small and the obtained scores are not fully accurate. However, these results show that the importance of the Δ parameters is at least as great as that of the MFCCs.

On the other side the score using the ΔΔ parameters set is not good (SER = 66.207 %) and suggests that ΔΔ parameters convey little useful information for recognition process.

The Δ parameters may be related to the instantaneous speed of change of the corresponding MFCC parameters evolution. In terms of trajectory, this speed may be characterized by the instantaneous slope of the trajectory. On the other hand, our proposed SSCF Angle parameters described in the previous chapters corresponds also to a way of characterizing the slope of a trajectory.

Secondly, the comparison of the two sets "MFCC + Δ + ΔΔ" and "Δ + ΔΔ" shows also that SER scores are similar, respectively 25.302 % and 26.572 %. The difference is just of 1.27 %, however, it is not very much because this ASR was considered in the small size of our training data. Reminding that the first vector is made up of 39 parameters, while the second one contains 26 parameters only. However, we can see that in our case, adding the values of the MFCC parameters does not improve much performance compared to a recognition system using only information on the instantaneous dynamics of the signal. This second comparison confirms also our hypothesis: the use of our new dynamic acoustic features, as an input of the recognition system, could provide possible improvements in performance. This is why we wish to check this hypothesis in the following second series of tests.

5.4 Our proposed dynamic acoustic speech features used as ASR's input

The results obtained in experiment 1 (section 5.3) show the main role of pseudo-dynamic parameters (Δ and ΔΔ of MFCCs) derived from MFCC parameters. In this section, we would like to focus on evaluation of the effect of SSCF Angles features in our Vietnamese ASR. In Chapter 4 above, SSCF Angles were proposed as dynamic acoustic speech features. While for MFCCs and their deviations, the dynamic approach is a static approach plus dynamic parameters taken into account, we can consider SSCF Angles as an intrinsically dynamic approach.

As presented in experiment 1 (section 5.3), MFCCs is the typical speech features currently used on automatic speech recognition system in clean environment with high performance. Therefore, in order to evaluate the effect of SSCF Angles in ASR, we consider the ASR results obtained from MFCCs and

their deviations as the baseline results to compare with the one obtained from SSCF Angles at the same Vietnamese ASR.

5.4.1 SSCF Angles computation in Vietnamese ASR

There is a difference in the computation of SSCF Angles in our ASR with the one as mentioned above in Chapter 4, it is difficult to determine the equivalent-transition sections in continuous speech. Therefore, we proposed to compute each SSCF Angle in Vietnamese ASR from SSCF parameters of two successive frames of speech that is described in Figure 5-2.

According to the analysis in section 4.1.2, we chose 6 sub-band filters; therefore, each speech frame will have 6 SSCF parameters.

And SSCF Angle is computed as in the following formula 5-1

$$\text{SSCF Angle(i)(i+1) (current)} = \text{atan}\left(\frac{\Delta SSCF_{i+1}}{\Delta SSCF_i}\right) \text{ in degree } (^o) \quad \textbf{(5-3)}$$

in which :

$\Delta SSCF_{i+1}$ = $SSCF_{i+1}$(current frame) – $SSCF_{i+1}$(current-1);

$\Delta SSCF_i$ = $SSCF_i$(current) – $SSCF_i$(current-1);

and i = 0, 1, 2, 3, 4, 5.

This computation of SSCF Angles is the simplest modeling and easy to deploy in an automatic speech recognition system.

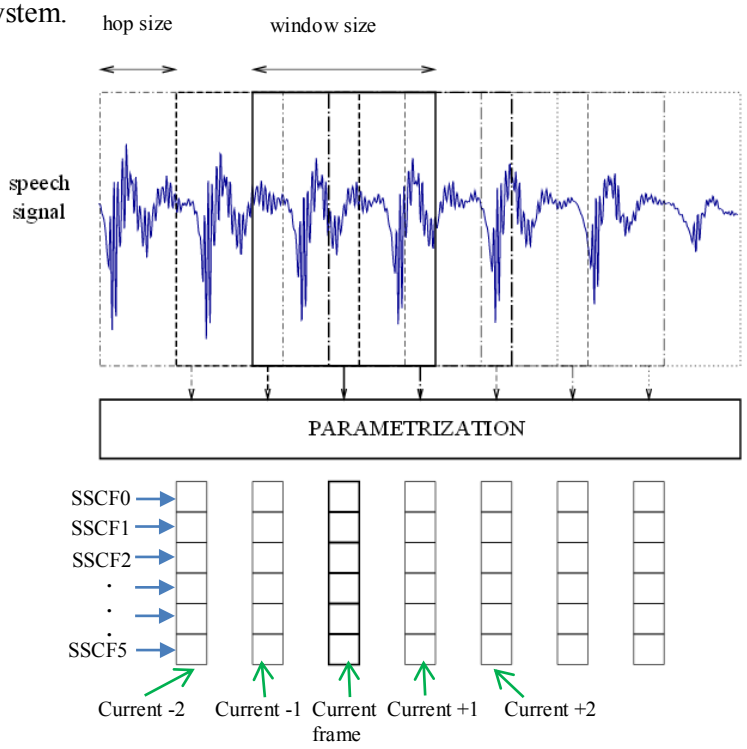


Figure 5-4: SSCF parameters extraction from a speech signal.

5.4.2 Experiment 2: ASR using SSCF Angles

5.4.2.1 Corpus

In Vietnamese, there are some differences in phonetics and tones depending on the different regional dialects (in the North, the South and the Centre of Vietnam) (Đoàn, 1999; Mark, 2002; Gregerson, 1969; Brunelle, 2009; Hoang, 1965; Michaud, 2004). In order to reduce the effect of dialect factors to ASR's performance, we select in this test a small part of our VietnameseCorpus including 12 Vietnamese native speakers from the North of Vietnam. The speech type is chosen common long paragraph. The corpus is divided to ensure the balance between male and female voices in both training and testing sets. This method is a popular way to perform on ASR in order to reduce the difference of speech features between male and female voices. Following to this approach, the corpus is divided as follows:

- Training set includes 3.9 speech hours from 8 Vietnamese speakers (4 males and 4 females).
- Testing set consist of 0.9 speech hours from 4 Vietnamese speakers (2 males and 2 females).

5.4.2.2 Method

As in experiment 1, we still use Sphinx toolkit to perform Vietnamese ASR on the corpus described above. Reminding that our ASR system only use an acoustic model (without language model).

As Vietnamese is a tonal language, we use tonal phoneme in Vietnamese ASR. For instance, in the sentence "Trời đẹp quá" [tɕɤ^{2}j^{2} dɛ^{6b}p^{6b} kw^{5}a^{5}], the following tonal phonemes were found: /tɕ/, /ɤ^{2}/, /j^{2}/, /d/, /ɛ^{6b}/, /p^{6b}/, /k/, /w^{5}/ and /a^{5}/. Each phoneme is modeled by a HMM model.

We will compare the results of SSCF Angles and/or Delta of SSCF Angles with the ones of MFCCs and/or delta and/or delta-delta at the same Vietnamese ASR. We call them ASR2.

In the same manner as in section 5.3, the ASR's quality is still evaluated by Syllable Error Rate (SER) factor in percent (%). The smaller SER the ASR system has, the better its quality is.

5.4.2.3 Results and discussions

The results on Vietnamese speech recognition of MFCCs, SSCF Angles and their derivations are presented in tables 5-2 and 5-3.

Table 5-2 shows that, as in the previous test (Section 5.3), the SER score obtained with a vector constituted only of Δ and ΔΔ parameters, (even in a smaller proportion), is closed to the SER obtained with the full vector (MFCC + Δ + ΔΔ). This suggests the importance of dynamic cues in recognition process and in coherent with our results obtained previously in Section 5.3.

Table 5-2: Syllable error rate (%) in Vietnamese ASR using MFCC and their derivation with the balance between male and female voices in both training and testing sets.

Types of parameters	SER (%)	Dimension of parameter vector
MFCC	29.050	13
Δ + ΔΔ	11.732	26
MFCC + Δ + ΔΔ	7.952	39

Table 5-3: Syllable error rate (%) in Vietnamese ASR using SSCF Angles and their derivation with the balance between male and female voices in both training and testing sets.

Types of parameters	SER (%)	Dimension of parameter vector
SSCF Angle	28.369	6
SSCF Angle + Δ	12.212	12

In these second series of experiments, it is interesting to compare the results obtained with the conventional MFCC parameters and those obtained with the SSCF Angle parameters (Tables 5-2 and 5-3).

For both ASR systems, if we use a vector of simple parameters, SER scores are similar: SER = 29.050 % with MFCC parameters and SER = 28.369 % with SSCF Angle parameters. However in the latter case, the number of parameters taken into account is half size (size of the SSCF Angle vector = 6, in comparison to the size of the MFCCs vector = 13). We can consider that we obtain similar results even if the signal characterization in the frequency domain contains less of useful information. It could be an advantage in terms of computational complexity and time.

If we add the first derivative (Δ) of the SSCF angle parameters, SER score is then similar to the case where only signal dynamics characteristics (Δ and ΔΔ) are computed from MFCCs parameters (respectively 12.212 % and 11.732 %). This result confirms our hypothesis according to which our way to characterize dynamic features produces similar results comparing to the classical approach.

However in this case the SER score (12.212 %) remains worse if we compare it to whose obtained by the classical input vector MFCC + Δ + ΔΔ (SER = 7.052 %). We can also add that for the use of the SSCF angle, the size of the input vector is more than 3 times smaller (12 in comparison to 39 for MFCC parameters and derivatives), which could be a bonus.

5.4.2.4 Conclusions

From our results, several conclusions can be drawn: firstly, it is clear for us that SSFC Angle parameters could be used in order to build an ASR system and these parameters allow the ASR system to produce similar results than one using classical MFCC parameters.

Secondly, the size of the input vector in the case of SSCF Angle parameters is more than 3 times smaller. We can conclude that, even if use less parameters than the classical MFCC parameters, these parameters contain enough useful information to be useable for an ASR system. It is an advantage in terms of computation cost.

Finally, however, SSFC Angle parameters are not sufficient by themselves and we need to add the first derivative Δ in order to obtain relatively good SER scores. This first derivative corresponds to the speed of the frequency transition. It is clear for us that the dynamic gesture must be characterized by the trajectory of the frequency transition (angle), plus by the speed of this transition.

5.4.3 Is our ASR system less dependent of speakers?

5.4.3.1 Our approach

In Chapter 4, we demonstrated that our proposed SSFC Angle parameters are fairly independent of the speakers for the same language: the measured values (to within the accuracy) are almost identical for all speakers, even if the speaker was a man or a woman. The existence of differences only depends on the language.

It is why we do the supposition that an ASR system using these SSFC Angle parameters as input vector could be intrinsically nondependent of the speakers.

In order to verify this hypothesis, we have done two new experiments to compare performances of our ASR system trained with MFCC parameters and those of our same ASR system trained with SSFC Angle parameters. For the first experiment we used only male speakers to train both versions of the ASR system and we used only female speakers to test them. In the second experiment we have done the opposite: female voices were used for the training set and male voices for the testing set. These corpora were called unbalanced corpus.

5.4.3.2 Corpus

We use a small part of VietnameseCorpus including Vietnamese native speakers from the North of Vietnam. The speech type is chosen common long paragraphs. This corpus is divided the unbalance between male and female voices in training and testing set.

For the first test, the training set includes 2.5 speech hours produced by 6 Vietnamese male speakers. The testing set consists in 0.5 speech hours from 2 Vietnamese female speakers.

For the second test, we have done the opposite: the training set includes 2.4 speech hours produced by 6 Vietnamese female speakers and the testing set consists in 0.45 speech hours from 2 Vietnamese male speakers.

As in the previous sections, both ASR system versions only use an acoustic model. We also evaluate the performance of ASR systems using Syllable Error Rate (SER) factor in percent (%).

5.4.3.3 Results and discussions

5.4.3.3.1 Experiment 3: First unbalanced test (training with males and test with females)

The Vietnamese ASR results on this test is presented in the tables 5-4 and 5-5, as follows:

Table 5-4: Syllable error rate (%) in Vietnamese ASR using MFCC and their derivations with the male training and female testing.

Types of parameters	SER (%)	Dimension of parameter vector
MFCC	74.505	13
Δ + ΔΔ	42.162	26
MFCC + Δ + ΔΔ	41.223	39

We firstly observe that the performances of the ASR system using MFCC parameters are not very good (see Table 5-4). They are even very bad if the system use a simple vector consisting of only the 13 MFCCs without using the derivatives Δ and ΔΔ: SER = 74.505 %. With the complete set of parameters performances just reach 41% for SER. This is not surprising and it is consistent with the results usually shown in the literature.

In the literature two approaches are typically used to make ASR systems less speaker-dependent. Either very large databases recorded with more than 100 speakers are used to train the ASR systems, or speaker adaptation or speaker normalization techniques are used. Speaker variation is one of the major error sources; consequently, in order to do speaker-independent speech recognition, the speaker variation effects can be minimized. To achieve this goal, multiple speaker clusters are constructed from the speaker-independent training database (Rabiner et al., 1979; Huang, 1992a). A codeword-dependent neural network is associated with each speaker cluster. The cluster that contains the largest number of speakers is designated as the golden cluster. This is intended to minimize distortions between acoustic data in each cluster and the golden speaker cluster, and the ASR performance was improved. In addition, if speaker-dependent data are available, the system could be normalized or adapted to a specific speaker such that the error rate could be significantly reduced. Since adaptation is based on speaker-independent parameter estimation criterion, and adapt those parameters that are less sensitive to the limited training data (Huang, 1991; 1992b).

Looking at the table 5-5, experiment 3 produces the same finding as those of the experience described in Section 5.3: using only single SSFC angle parameters is not sufficient to achieve acceptable performances and it is necessary to add the derivative Δ (i.e. the speed of the frequency

transition) to the input vector. With this set of parameters (SSCF Angle + Δ), we obtained a SER equal to 25.444 %.

Table 5-5: Syllable error rate (%) in Vietnamese ASR using SSCF Angles and their derivations with the male training and female testing.

Types of parameters	SER (%)	Dimension of parameter vector
SSCF Angle	48.907	6
SSCF Angle + Δ	25.444	12

On the other hand, a comparison between the best results obtained by the ASR system using MFCC parameters (MFCCs + Δ + ΔΔ→ SER = 41.223 %, in Table 5-4) and those obtained by the ASR version using SSFC Angle parameters shows that the system using the latter is much better (almost two times smaller: SER = 25.444 %). In addition, the fact that in the first case the input vector contains 39 parameters while in the second case it contains only 12 parameters, demonstrates that the second system which uses SSCF Angle + Δ seems doubtless better.

This observation suggests to us that the dynamic acoustic speech features (SSCF Angles) still perform much better than the acoustic static speech feature (MFCCs) and/or even the combination of MFCCs with its derived dynamic features and, in fact, help to build an ASR system more independent of the speakers.

5.4.3.3.2 Experiment 4: Second unbalanced test (training with females and testing with females)

The Vietnamese ASR results on this test is presented in the table 5-6 and 5-7, as follows:

Table 5-6: Syllable error rate (%) in Vietnamese ASR using MFCC and their derivations with the female training and male testing.

Types of parameters	SER (%)	Dimension of parameter vector
MFCC	92.071	13
Δ + ΔΔ	54.179	26
MFCC + Δ + ΔΔ	50.536	39

Table 5-6 shows that similar results are found in this test as those of the experience 3 described in the previous section: (i) ASR system using simple 13 MFCC parameters without Δ and ΔΔ is very bad (SER = 92.071 %); (ii) With the complete set of parameters, performances just reach the 50.536 % for SER, in which, SER = 54.179 % for ASR using the set of Δ and ΔΔ; (iii) therefore, in this test, the derivatives play the main role in ASR.

Table 5-7: Syllable error rate (%) in Vietnamese ASR using SSCF Angles and their derivations with the male training and female testing.

Types of parameters	SER (%)	Dimension of parameter vector
SSCF Angle	42.875	6
SSCF Angle + Δ	31.911	12

Besides, a comparison between the best results obtained by the ASR system based on MFCC parameters (MFCCs + Δ + ΔΔ→ SER = 50.536 %, see Table 5-6) and those obtained by the ASR version using SSFC Angle parameters shows that the system using the SSCF Angle is much better. In the latter case, the SER is almost one point five times smaller (SER = 31.911 %). As in the first case the input vector contains 39 parameters while in the second case it contains only 12 parameters, we can say that the second system using SSCF Angle + Δ seems much better.

Combining with the results obtained in experiment 3, these observations suggests once again that the dynamic acoustic speech features (SSCF Angles) still perform much better than the acoustic static speech feature (MFCCs) and/or even the combination of MFCCs with their derived dynamic features. Consequently, they contribute to build an ASR system more independent of the speakers.

5.5 Discussion

These simple but encouraging first tests of an automatic speech recognition allow us to conclude that our proposed SSFC Angle parameters can be used directly as input parameters of an automatic speech recognition system, with similar results to those obtained with a system that conventionally uses MFCCs as input parameters.

We have even shown that, for equivalent recognition scores, our SSCF Angle parameters are better because they constitute input vectors which have a size 2 or 3 times smaller. This point is an indisputable gain in terms of complexity of calculations for recognition models.

We have also shown that our recognition system using our SSCF Angle parameters is much more independent of the speakers.

However, although their performance is much better, they seem not optimal so far. A first direction further study is to improve the computation of the SSCF Angle parameters used in the speech recognition system. We implemented this calculation by a simple measurement of angle between two consecutive frames (as explained in Section 5.4.1). Each frame corresponds to a time window of 20ms only. However, in our theoretical study, the angles were calculated on a much larger portion of the vowel-vowel transition. Therefore, in order to obtain better input parameters for the ASR system, it might be wiser to calculate parameters on more frames, 3 or 5 for example. So our parameters would

probably be less sensitive to instantaneous changes in the speech signal and thus closer to the theoretical values measured in Chapter 4.

Another question concerns the evolution of these angle features for vowels-consonants transitions. During our study presented in Chapter 4, we characterized these angle parameters only for simple vowel-vowel transitions. We have not characterized measurements during the production of consonants. It should be checked so that during this consonant production, our characterization of the acoustic gesture remains valid. Will a single measurement of angle always be relevant? Does this single measure remain valid for the consonant gesture? Or will it be necessary to decouple between the vowel gesture and the consonant gesture and thus to add a second parameter which then characterizes the consonant gesture?

Moreover, in these first tests of speech recognition, we integrated the frequency measurement SSCF0, and also calculated its transition angle. However, all our theoretical study did not address this SSCF0 parameter because we made the assumption that it essentially represents the frequency band corresponding to the fundamental frequency F0. However, while this approximation can be considered acceptable for open vowels, this is no longer true for the vowels which present a very low F1: SSCF0 then represents a mixture of F0 and F1.

Finally, a further improvement can be done when we take into account the transition speed. We found it is necessary to compute it, which seems natural enough: when one wants to characterize a "gesture", one needs to specific at least its trajectory and also its speed. We just simply calculated this speed by the first derivative of the angle. It would probably be much better to characterize this speed also by an angle measurement on the speed trajectory itself. Indeed, if we take the figure 4.13 presented in Section 4.2.1.2, we can see that the speed has also trajectories that may, at first approximation, be considered as a straight line. An angle measurement is also possible and simple. We could then have, as the input vector of the recognition system, a series composed of 6 SSCF Angles plus 6 SSCF Speed angl

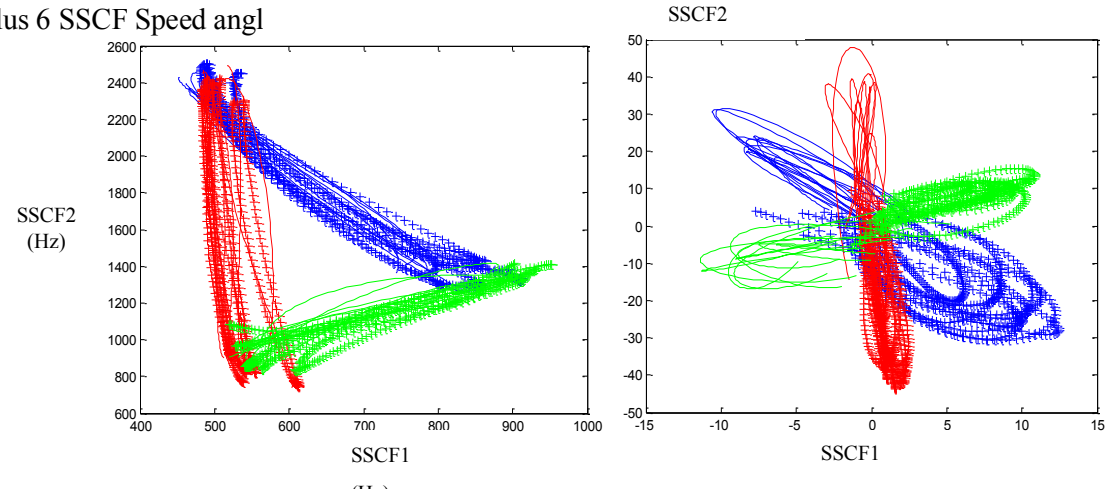


Figure 5-5 (from Figure 4-13): Vowel-to-Vowel transitions on SSCF1/SSCF2 plane and transition speeds produced by a native male speaker of Vietnamese.

5.6 Conclusions of chapter 5

In Chapter 5, we built a simple but classical recognition engine for Vietnamese language and we carried out two ASR versions in order to conduct comparisons: (i) a first version using MFCC parameters for the characterization of the speech signals, parameters usually used in many works of literature, and (ii) a second version that allows us to test our new parameters based on the angles of the transitions of our spectral parameters SSCFs.

During the first series of tests, we verified that the dynamic aspects are important for speech recognition process. It was a well-known point of literature, but we wanted to check it with our speech recognition system. We have thus verified that the two derivatives (the speed and acceleration parameters), conventionally used to take into account of the dynamic aspects of the speech signal, carry more useful information in the recognition process than those using single MFCC parameters.

Then, in the second series of tests, we have shown that it is possible to use our angle parameters as input vector of our speech recognition system, and the obtained result sets were quite similar to those of a classical recognition system.

During the third series of tests, we achieved recognition tests with unbalanced gender databases. The results showed that our ASR system using the SSCF angles is much less dependent on the speakers than those using MFCCs. This confirms our hypothesis that the characterization of dynamic acoustic gestures can be a great advantage for automatic speech recognition because it allows to design speech recognition systems that are intrinsically independent of the speakers.

However, these first tests, even if they are very interesting, are not enough. They showed that we have to improve our recognition system. The first possible directions for improvements are: (i) a better calculation of angles, closer to the theory, in order to obtain an input vector more representative of the acoustic gesture; (ii) a better characterization of these acoustic gestures during production of consonants; (iii) taking better account of the speed of the transition, not only in calculating the derivative of the angle, but in calculating a specific “speed” angle, directly measured on speed transitions.

Finally, we are aware that we must do our tests with a much larger training corpus which will be recorded by more speakers. The results will be more reliable and will then allow better comparison.

Chapter 6
Conclusions and Perspectives

6.1 Conclusions

Dynamic acoustic speech features throughout are the guiding thread running through this entire dissertation, in which they were studied in speech perception, speech analysis, and speech recognition for Vietnamese.

Chapter 1 on state of the art of speech feature expanded on the many limitations of a static view of speech. A wealth of research in the second half of the twentieth century demonstrated that speech is a dynamic process in both speech production and speech perception, these works opened a range of new paths and strands of research, to which this dissertation aimed to make a further contribution. Adopting the perspective that dynamics are of the essence when dealing with speech, our thesis work from Chapter 2 to Chapter 5 proposed some acoustic dynamic speech features, providing a degree of detail as to their analysis and their potential applications in recognition system, with Vietnamese – for obvious reasons – as a first testing-ground.

Chapter 2 studied some speech characteristics on both static and dynamic aspects, both in the speech signal and at the perception level. Two experiments were carried out to test the effects of impulse at the beginning, at the end and during transition of the speech signals. The obtained results showed that these effects are slight – which comes as good news in the perspective of a search for stable acoustic correlates of speech segments. Besides, speech signal is completely characterized by its power spectrum and phase spectrum. Therefore, we moved on to study the role of phase spectrum which is considered next dynamic characteristic candidate of speech in perception test. The perceptual result showed that the amplitudes of spectral components are clearly sufficient for perspective discrimination without information of phase spectrum.

In view of up the results obtained in Chapter 2, the work of next chapters considered in finding out another dynamic candidate of speech signal (not phase spectrum) on amplitude domain that can be useful for speech discrimination.

In Chapter 3, we studied the role of formant transitions of speech perception V1CV2. The obtained results showed that perception of synthesized V1CV2 can be obtained with formant frequency transition situated both inside and outside the vowel triangle. These results extended the previous results published in the deductive approach of Carré (2004) proposing a dynamic view of speech production, and on the prediction of vocalic systems of Carré (2007, 2009a, 2009b) and Nguyen (2009): transition direction and length on frequency domain, and of transition rate (slope) in time domain can distinguish VV, CV stimuli. However, all these parameters are based on formant frequency transition, so that they suffer from inherent limitations in estimating the formant frequencies, therefore our work in Chapter 4 aimed to find out another speech feature as pseudo-formant frequencies. These parameters are Spectral Subband Centroid Frequency (SSCF) features. Some analysis works was performed to show that SSCF parameters are similar to formant frequencies but they are continuous even during consonant production contrary to formant frequencies.

Based on SSCF parameters, we proposed a new way to modeling the dynamic speech features and we called them SSCF Angles.

Some analysis work on SSCF Angles parameters were performed on some transitions of Vowel-to-Vowel (V1V2) of both Vietnamese and French. SSCF Angles are precise and reliable parameter. The analysis results show that SSCF Angles allow to distinguish the V1V2 transitions. The symmetric property of V1V2 and V2V1 on acoustic domain based on SSCF Angles.

The analysis results on SSCF Angles also showed that these parameters are fairly independent for male and female on the same transition VV study context. And they are also fairly invariant for speech rate (normal and fast speech rates).

Keeping up the interesting analysis results of SSCF Angles in Chapter 4, we applied these parameters into automatic Vietnamese speech recognition systems in order to test its performance and then give the comparison with MFCCs – the most popular speech feature currently used on automatic speech recognition system in clean environment with high performance. MFCCs include the static parameters MFCC and/or its derived dynamic features (delta and/or delta-delta). Some interesting ASR results we obtained, are presented in Chapter 5.

During the first series of tests, we verified that the dynamic aspects are important for speech recognition process. We have thus verified that the two derivatives (the speed and acceleration parameters), conventionally used to take into account dynamic aspects of speech signal, carry more useful information in the recognition process than those using single MFCC parameters.

Then, in the second series of tests, we have shown that it is possible to use our SSCF angle parameters as an input vector of our speech recognition system, and the obtained result sets were quite similar to those of a classical recognition system.

During the third series of tests, we achieved recognition tests with unbalanced gender databases. The results showed that our ASR system using the SSCF angles is much less dependent on the speakers than those using MFCCs. This confirms our hypothesis that the characterization of dynamic acoustic gestures can be a great advantage for automatic speech recognition because it allows for the design of speech recognition systems that are intrinsically independent of the speakers.

6.2 Perspectives

From the obtained results, our proposed acoustic dynamic speech feature SSCF Angles proved that they are useful parameters on automatic Vietnamese speech recognition system, especially when the speech database set is unbalanced. On the other hand, while the obtained results appear promising, the ASR results based on unbalanced corpus are not very good. This is a clear indication that we have to continue studying for improving ASR's performance.

Firstly, we will have to do our tests with a much larger training corpus which will be recorded by more speakers so that the results will be more reliable and will then allow better comparison.

Secondly, improving the computation way of the SSCF Angle parameters used in the speech recognition system is actually necessary. We implemented this calculation by a simple measurement of angle between two consecutive frames. Each frame corresponds to a time window of 20ms only. However, in our theoretical study, the angles were calculated on a much larger portion of the vowel-vowel transition. Therefore, in order to obtain better input parameters for the ASR system, it is necessary to calculate SSCF Angle parameters on more frames, 3 or 5 for example and do devise ways to test these parameters in ASR, in the same way as in the experiments 2 to 4set out in Chapter 5 of this dissertation.

Thirdly, a further improvement can be done when we take into account the transition speed. We found that it is necessary to compute it, which seems normal when one wants to characterize a "gesture" which must be at least characterized by its trajectory and also its speed. In our thesis study, we just simply calculated this speed by the first derivative of the angle. Therefore, it would probably be much better to characterize this speed also by an angle measurement on the speed trajectory itself.

Fourthly, we characterized these angle parameters only for simple vowel-vowel transitions. We have not characterized measurements during the production of consonants. This is a further point that calls for verification, so that during this consonant production, our characterization of the acoustic gesture remains valid. From that point, we will find out the way to characterize consonant gesture appropriately. On this basis, one can consider the possibilities for using those parameters: combining the SSCF Angle parameters in an automatic speech recognition system.

According the same above approach, it is necessary for us to enlarge the dynamic acoustic analysis for all Vietnamese transitions of Vowel-Vowel, Consonant-Vowel-Consonant and Vowel-Consonant

so that we can find out the best-case way to characterize the dynamic acoustic speech features on whole continuous speech.

Our study in this dissertation just considered the analysis on male and female adult voices. So in the future work, analyzing on child voices will be needed in order to have an overall view on dynamic acoustic speech features along with their role in ASR.

All experiments on ASR in our thesis were performed with relatively 'clean' speech recorded in a sound-treated studio. This raises the issue of what happens with our proposed speech features about the case of noise signals. Thus, we intend to develop further studies with such signals in future work. This will provide a further opportunity to put to the test the robustness of the proposed acoustic features.

Finally, in order to confirm the usefulness of those dynamic acoustic speech features in both speech perception and speech recognition system, more studies will be performed on other languages with the same above approach as here on Vietnamese.

Bibliography

Agwuele, A., Sussman, H.M., Lindblom, B., 2008. The effect of speaking rate on consonant vowel coarticulation. Phonetica 65, 194–209. doi:10.1159/000192792

Alliot, A., 2009. Reconnaissance de la parole par modélisation des gestes. Stage de fin d'études – Mica, Vietnam.

Alsteris, L.D., Paliwal, K.K., 2007. Short-time phase spectrum in speech processing: A review and some experimental results. Digit. Signal Process. 17, 578–616. doi:10.1016/j.dsp.2006.06.007

Atal, B.S., Schroeder, M.R., 1974. Recent advances in predictive coding - applications to speech synthesis. Prepr. Speech Commun. Semin. Stockh.

Atlas, L., 2006. Modulation frequency filtering of speech, in: Dynamics of Speech Production and Perception, NATO Science Series Series I. Life and Behavioural Sciences. Ios Press, Amsterdam ; Washington, DC, pp. 195–205.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C., 2007. Automatic speech recognition and speech variability: A review. Speech Commun. 49, 763–786. doi:10.1016/j.specom.2007.02.006

Brancazio, L., Fowler, C.A., 1998. On the relevance of locus equations for production and perception of stop consonants. Percept. Psychophys. 60, 24–50.

Brunelle, M., 2009. Tone perception in Northern and Southern Vietnamese. J. Phon. 37, 79–96. doi:10.1016/j.wocn.2008.09.003

Calliope Firm, Tubach, J.P., Fant, G., 1989. La Parole et son traitement automatique. Masson, Paris; Milan; Barcelone.

Carré, R., 2009a. Signal dynamics in the production and perception of vowels. Approaches Phonol. Complex. F PellegrinoE Marsico Chitoran C Coupé Eds Mouton Gruyter Berl. N. Y. 59–81.

Carré, R., 2009b. Dynamic properties of an acoustic tube: Prediction of vowel systems. Speech Commun. 51, 26–41. doi:10.1016/j.specom.2008.05.015

Carré, R., 2008. Production and perception of V1V2 described in terms of formant transition rates. J. Acoust. Soc. Am. 123, 3324. doi:10.1121/1.2933811

Carré, R., 2004. From an acoustic tube to speech production. Speech Commun. 42, 227–240. doi:10.1016/j.specom.2003.12.001

Carré, R., Hombert, J.M., 2002. Variabilité phonétique en production et perception de parole : stratégies individuelles (Phonetic variabilities in speech production and perception: individual strategies). Invariants Var. Dans Sci. Cogn. Mazoyer B Lautrey J Van Geert P Eds Paris Press. Maison Sci. Homme.

Carré, R., Lancia, R., 1975. Perception of vowel amplitude transients. edited by M. A. A. Tatham (Academic Press, London).

Carré, R., Mrayati, M., 1991. Vowel-vowel trajectories and region modeling. J Pho- Netics 19 433–443.

Carré, R., Pellegrino, P., Divenyi, P., 2007. Speech dynamics: epistemological aspects. Proc 16th Int. Congr. Phon. Sci. ICPhS.

Carré, R., Quach, T.N., 1987. Effects of nonstationary characteristics on vowel perception. Bull. Lab. Commun. Parlée 1 307–318.

Castelli, E., 2013. Vocalic space normalisation (Internal report). International Institute MICA.

Castelli, E., Carré, R., 2005. Production and perception of Vietnamese vowels. INTERSPEECH 2005 - Eurospeech 9th Eur. Conf. Speech Commun. Technol. Lisbon Port. 2881–2884.

Chiba, T., Kajiyama, M., 1958. The vowel, its nature and structure. Phonetic Society of Japan.

Christovich, L., Kozhevnikov, V., 1970. Theory and methods on the perception of speech signals. Washington, D, C.; National Technical Information Service, U.S. Department of Commerce.

CMU Sphinx, 2015. http://cmusphinx.sourceforge.net/. Carnegie Mellon Univ.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. Vol 28 No 4 357–366.

De Cheveigné, A., 2003. Time-domain auditory processing of speech. J. Phon. 31, 547–561. doi:10.1016/S0095-4470(03)00041-X

Divenyi, P., Greenberg, S., Meyer, G. (Eds.), 2006. Dynamics of speech production and perception, NATO science series Series I. Life and behavioural sciences. Ios Press, Amsterdam ; Washington, DC.

Divenyi, P.L., 2005. Frequency change velocity and acceleration detector: A bird or a red herring?, in: Pressnitzer, D., de Cheveigné, A., McAdams, S., Collet, L. (Eds.), Auditory Signal Processing. Springer New York, New York, NY, pp. 176–184.

Đoàn, T.T., 1999. Ngữ âm tiếng Việt (translate: Vietnamese phonology). Hanoi national university publishing house.

El-Samie, F.E.A., 2011. Information security for automated speech identification, Springer Briefs in speech technology. Springer, New York.

Fant, G., 1960. Acoustic theory of speech production. The Hague: Mouton.

Fant, G., Martony, J., 1962. Speech synthesis instrumentation for parametric synthesis (OVE II). Peech Transm. Lab. Q. Prog. Status Rep. 3, 18–24.

Fowler, C.A., 1980. Coarticulation and theories of extrinsic timing. J Phon. 8 113–133.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoust. Speech Signal Process. 34, 52–59. doi:10.1109/TASSP.1986.1164788

Gay, T., 1978. Effect of speaking rate on vowel formant movements. J. Acoust. Soc. Am. 63 223–230.

Gendrot, C., Adda-Decker, M., 2007. Impact of duration and vowel inventory size on formant values of oral vowels: an automated formant analysis from eight languages. Int. Conf. Phon. Sci. 1417–1420.

Glass, J., 2007. A brief introduction to automatic speech recognition (Slideshow of course taught at MIT Computer Science and Artificial Interlligence Laboratory).

Graetzer, S., 2008. Coarticulation in CV sequences: Locus equation data. J. Acoust. Soc. Am. 123, 3328. doi:10.1121/1.2933830

Greenberg, S., Carvey, H., Hitchcock, L., Chang, S., 2002. Beyond the phoneme: a juncture-accent model of spoken language. Proc. Second Int. Conf. Hum. Lang. Tecnol. Res. Morgan Kaufmann Publ. Inc San Franc. CA USA 36–43.

Gregerson, K., 1969. A study of Middle Vietnamese phonology. Bull. Société Etudes Indochinoises 44 135–193.

Guenther, F.H., 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychol. Rev. 102, 594–621.

Hermansky, H., 2011. Speech recognition from spectral dynamics. Sadhana 36, 729–744. doi:10.1007/s12046-011-0044-2

Hermansky, H., 1994. RASTA processing of speech. IEEE Trans Speech Audio Proc Vol 2 No 4 578–589.

Hinton, G., Deng, L., Yu, D., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Dahl, T.S.G., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. 82–97.

Hirtle, D., 2004. Speech variability: the biggest hurdle for recognition. Intern. Rep. Fac. Comput. Sci. Univ. N. B.

Hoang, T.C., 1965. Tiếng Việt trên các miền đất nước: Phương ngữ học (translate: Vietnamese in the country: The linguist). NXB. Khoa Hoc Xa Hoi, Hanoi, Vietnam.

Huang, X., 1992a. Minimizing speaker variation effects for speaker-independent speech recognition. Association for Computational Linguistics, p. 191. doi:10.3115/1075527.1075569

Huang, X., 1992b. Speaker normalization for speech recognition. IEEE, pp. 465–468 vols.1. doi:10.1109/ICASSP.1992.225871

Huang, X., 1991. A study on speaker-adaptive speech recognition. Association for Computational Linguistics, pp. 278–283. doi:10.3115/112405.112458

Huang, X., Acero, A., Hon, H.-W., 2001. Spoken language processing: a guide to theory, algorithm, and system development. Prentice Hall PTR, Upper Saddle River, NJ.

Huggins, W.H., 1952. A phase principle for complex frequency analysis and its implications in auditory theory. J Acoust Soc Am Vol 24 582–589.

Hunt, M.J., 1987. Delayed decisions in speech recognition - The case of formants. Pattern Recognit. Lett. 6, 121–137. doi:10.1016/0167-8655(87)90093-6

Jelinek, F., 1976. Continuous speech recognition by statistical methods. Proc. IEEE 64, 532–556. doi:10.1109/PROC.1976.10159

Johnson, K., 1997. Speaker perception without speaker normalization. An exemplar model. K Johns. J W Mullennix Eds Talker Var. Speech Process. Acad. Press N. Y.

Johnson, K., 1990. Contrast and normalization in vowel perception. J Acoust Soc Am 24.

Johnson, K., Flemming, E., Wright, R., 1993. The hyperspace effect: Phonetic targets are hyperarticulated. Language 69, 505. doi:10.2307/416697

Kandel, E.R., Schwartz, J.H., Jessell, T.M., 2000. Principles of neural science, 4th ed. ed. McGraw-Hill, Health Professions Division, New York.

Keating, P.A., 1990. The window model of coarticulation: Articulatory evidence. Kingston J Beckman M E Eds 451–470.

Kent, R.D., Moll, K.L., 1969. Vocal tract characteristics of the stop cognates. J. Acoust. Soc. Am. 1549–1555.

Klatt, D.H., 1977. Review of the ARPA speech understanding project. J. Acoust. Soc. Am. 62, 1345–1366. doi:10.1121/1.381666

Ladefoged, P., Disner, S.F., 2012. Vowels and consonants, 3rd ed. ed. Wiley-Blackwell, Malden, MA.

Laprie, Y., Berger, M., 1992. Active models for regularizing formant trajectories. Second Int. Conf. Spok. Lang. Process. ICSLP 1992.

Le, V.B., Tran, D.D., Castelli, E., Besacier, L., Serignat, J.-F., 2004. Spoken and written language resources for Vietnamese. Acts LREC 599–602.

Liberman, A.M., Delattre, P.C., Cooper, F.S., Gerstman, L.J., 1954. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monogr. Gen. Appl. 68, 1–13. doi:10.1037/h0093673

Lindblom, B., 2004. The organization of speech movements: specification of units and modes of control. Sound Sense Workshop MIT.

Lindblom, B., 1963a. Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 783–783. doi:10.1121/1.2142410

Lindblom, B., 1963b. On vowel reduction. R. Inst. Technol. Speech Transm. Lab. Stockh.

Lindblom, B., Mauk, C., Moon, S.-J., 2006. Dynamic specification in the production of speech and sign, in: Dynamics of Speech Production and Perception, NATO Science Series, I: Life and Behavioural Sciences. pp. 7–20.

Lindblom, B., Sussman, H.M., 2012. Dissecting coarticulation: How locus equations happen. J. Phon. 40, 1–19. doi:10.1016/j.wocn.2011.09.005

Lotto, A.J., Kluender, K.R., Holt, L.L., 1997. Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). J. Acoust. Soc. Am. 102, 1134–1140.

Lublinskaja, V.V., Ross, J., Ogorodnikova, E.V., 2006. Auditory perception and processing of amplitude modulation in speech-like signals. Legacy of the Chistovich - Kozhevnikov Group.

Mai, N.C., Vu, D.N., Hoang, T.P., 1997. Cơ sở ngôn ngữ học và tiếng Việt (Tr. Les bases linguistiques et la langue vietnamienne). Hanoi: Nhà xuất bản Đại học và Giáo dục chuyên nghiệp (Edition de l'enseignement supérieur et de l'éducation professionnelle).

Mark, A., 2002. A look at North-Central Vietnamese. SEALS XII 12th Annu. Meet. Southeast Asian Linguist. Soc.

Michaud, A., 2005. Prosodie de langues à tons (naxi et vietnamien), prosodie de l'anglais : éclairages croisés (In: Cahiers de linguistique - Asie orientale). Paris: Université Paris 3 - Sorbonne Nouvelle.

Michaud, A., 2004. Final consonants and glottalization: New perspectives from Hanoi Vietnamese. Phonetica 61, 119–146. doi:10.1159/000082560

Monsen, R.B., Engebretson, A.M., 1983. The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction. J. Speech Hear. Res. 26, 89–97.

Moon, J.S., Lindblom, B., 1994. Interaction between duration, context and speaking style in English stressed vowels. J Acoust Soc Am 96 40–55.

Nathan, G.S., Cao, X.H., 1988. Phonologie et linéarité: Réflexions critiques sur les postulats de la phonologie contemporaine. Language 64, 193. doi:10.2307/414808

Neel, A.T., 2008. Vowel space characteristics and vowel identification accuracy. Ournal Speech Lang. Hear. Res. 51, 574–585.

Newell, A., 1973. Speech understanding systems: final report of a study group. North-Holland Publ. Co, Amsterdam.

Nguyen, T.T.H., 2004. Contribution à l'étude de la prosodie du Vietnamien (Variations de l'intonation dans les modalités - assertive, interrogative et impérative). Paris 7: Université Denis Diderot.

Nguyen, V.S., 2009. Etude de caracteristiques de la langue vietnamienne en vue de sa synthese et de sa reconnaisance automatique aspects statiques et dynamiques.

Nguyen, V.S., Carré, R., Castelli, E., 2008. Locus equation for final stop voiceless consonants /p, t, k/ in Vietnamese language. Proc. Empir. Methods Asian Lang. Process. Workshop 123–132.

Nguyen, V.S., Castelli, E., Carré, R., 2010. Production and perception of Vietnamese final stop consonants /p, t, k/. Second Int. Workshop Spok. Lang. Technol. -Resour. Lang. SLTU10 136–141.

Nguyen, V.S., Castelli, E., Carré, R., 2009. Vietnamese final stop consonants /p, t, k/ described in terms of formant transition slopes. IEEE, pp. 86–90. doi:10.1109/IALP.2009.27

Nordström, P.E., Lindblom, B., 1975. A normalization procedure for vowel formant data. 8th Int. Congr. Phon. Sci. Leeds.

Paliwal, K.K., 1998. Spectral subband centroids as features for speech recognition. IEEE, pp. 124–131. doi:10.1109/ASRU.1997.658996

Paliwal, K.K., Alsteris, L.D., 2003. Usefulness of phase spectrum in human speech perception. N Eurospeech 2003 Geneva.

Perkell, J.S., Matthies, M.L., Svirsky, M.A., Jordan, M.I., 1993. Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: a pilot "motor equivalence" study. J. Acoust. Soc. Am. 93, 2948–2961.

Peterson, G.E., Barney, H.L., 1952. Control methods used in the study of the vowels. J Acoust Soc Am 24 175–184.

Pisoni, D.B., Remez, R.E., 2005. The handbook of speech perception. Blackwell Publishing Ltd, Oxford, UK.

Pollack, I., 1968. Detection of rate of change of auditory frequency. J. Exp. Psychol. 77(4), 535–541.

Pols, L.C.W., Van Son, R.J.J., 2006. Speech dynamics: acoustic manifestations and perceptual consequences, in: Dynamics of Speech Production and Perception, NATO Science Series, I: Life and Behavioural Sciences. pp. 71–80.

Rabiner, L., Levinson, S., Rosenberg, A., Wilpon, J., 1979. Speaker independent recognition of isolated words using clustering techniques. Institute of Electrical and Electronics Engineers, pp. 574–577. doi:10.1109/ICASSP.1979.1170821

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of speech recognition, Prentice Hall signal processing series. PTR Prentice Hall, Englewood Cliffs, N.J.

Robert, M., 2009. Introduction to Speech Production (http://clas.mq.edu.au/speech/phonetics/phonetics/introduction). Macquarie Univ. Syney Aust.

Saksamudre, S.K., Shrishrimal, P., Deshmukh, R., 2015. A review on different approaches for speech recognition system. Int. J. Comput. Appl. 115.

Schmid, P., Barnard, E., 1995. Robust, n-best formant tracking. Fourth Eur. Conf. Speech Commun. Technol. EUROSPEECH 1995.

Serniclaes, W., 1987. Etude expérimentale de la perception du trait de voisement des occlusives du français Unpublished Ph. D. thesis, Université Libre de Bruxelles. Université Libre de Bruxelles.

Serniclaes, W., Sprenger-Charolles, L., Carre´, R., Demonet, J.-F., 2001. Perceptual discrimination of speech sounds in developmental dyslexia. J. Speech Lang. Hear. Res. 44, 384. doi:10.1044/1092-4388(2001/032)

Shankweiler, D., Verbrugge, R.R., Studdert-Kennedy, M., 1978. Insufficiency of the target for vowel perception. J Acoust Soc Am 63 S4.

Stern, R.M., Trahiotis, C., Ripepi, A.M., 2006. Fluctuations in amplitude and frequency enable interaural delays to foster the identification of speech-like stimuli, in: Dynamics of Speech Production and Perception, NATO Science Series, I: Life and Behavioural Sciences. pp. 143–151.

Stetson, R.H., 1928. Motor phonetics. Archives neerlandaises de Phonetique Experimental.

Steve Young, 1996. A review of large-vocabulary continuous-speech. IEEE Signal Process. Mag. 13, 45. doi:10.1109/79.536824

Stevens, K.N., 2000. Acoustic phonetics. Current studies in linguistics. MIT Press, Cambridge, Mass.

Stevens, K.N., House, A.S., 1963. Perturbation of vowel articulations by consonantal context: An acoustical study. J. Speech Lang. Hear. Res. 6, 111. doi:10.1044/jshr.0602.111

Strange, W., 1989a. Dynamic specification of coarticulated vowels spoken in sentence context. J. Acoust. Soc. Am. 85, 2135–2153.

Strange, W., 1989b. Evolving theories of vowel perception. J. Acoust. Soc. Am. 85, 2081–2087.

Strange, W., Jenkins, J.J., Johnson, T.L., 1983. Dynamic specification of coarticulated vowel. J. Acoust. Soc. Am. 695–705.

Sussman, H.M., McCaffrey, H.A., Matthews, S.A., 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. J. Acoust. Soc. Am.

Sussman, H.M., Shore, J., 1996. Locus equations as phonetic descriptors of consonantal place of articulation. Percept. Psychophys. 58, 936–946.

Sweet, H., 1877. A handbook of phonetics. Clarendon Press, College Park, MD: McGrath.

Taylor, P.A., 2009. Text-to-speech synthesis. Cambridge University Press, Cambridge, UK ; New York.

Trần, T.T.H., 2011. Processus d'acquisition des clusters et autres séquences de consonnes en langue seconde : de l'analyse acoustico-perceptive des séquences consonantiques du vietnamien à l'analyse de la perception et production des clusters du français par des apprenants vietnamiens du FLE. Université Stendhal - Grenoble 3.

Trương, V.C., 1970. Structure de la langue vietnamienne. Paris : Imprimerie nationale : P. Geuthner.

Turner, R.E., Al-hames, M.A., Smith, D.R.R., Kawahara, H., Irino, T., Patterson, R.D., 2006. Vowel normalization: time-domain processing of the internal dynamics of speech, in: Dynamics of Speech Production and Perception, NATO Science Series, I: Life and Behavioural Sciences. p. 153–170.

Vaissière, J., 2011. On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. Proc. ICPhS XVII 52–59.

Vallée, N., 1994. Systèmes vocaliques : de la typologie aux prédictions. Université Stendhal Grenoble.

Verbrugge, R.., Rakerd, B., 1980. Talker-independent information for vowel identity. Haskins Lab. Status Rep. Speech Res. SR-62 205–215.

Vicsi, K., 2006. Computer-assisted pronounciation teaching and training methods based on the dynamic spectro-temporal characteristics of speech, in: Dynamics of Speech Production and Perception, NATO Science Series, I: Life and Behavioural Sciences. pp. 283–304.

Vilain, C., Berthommier, F., Boe, L.-J., 2015. A brief history of articulatory-acoustic vowel representation. 1st Int. Workshop Hist. Speech Commu- Nication Res. HSCR 2015.

Website 1, 2007. http://home.cc.umanitoba.ca/~robh/archives/arc0701.html.

Website 2, 2007. http://home.cc.umanitoba.ca/~robh/archives/arc0702.html.

Website 3, 2009. http://home.cc.umanitoba.ca/~robh/howto.html.

Weenink, D.J.., 2015. Improved formant frequency measurements of short segments. 18th Int. Congr. Phon. Sci.

Welling, L., Ney, H., 1996. A model for efficient formant estimation. IEEE, pp. 797–800. doi:10.1109/ICASSP.1996.543241

Whalen, D.H., Magen, H.S., Pouplier, M., Min Kang, A., Iskarous, K., 2004. Vowel production and perception: Hyperarticulation without a hyperspace effect. Lang. Speech 47, 155–174. doi:10.1177/00238309040470020301

Wiki-English, 2015. https://en.wikipedia.org/wiki/Vietnamese_language.

Wiki-French, 2015. https://en.wikipedia.org/wiki/French_language.

Wood, S., 1989. The precision of formant frequency measurement from spectrograms and by linear prediction. STL-QPSR 30, 091–094.

Huang, X., Acero, A., Hon, H.-W., 2001. Spoken language processing: a guide to theory, algorithm, and system development. Prentice Hall PTR, Upper Saddle River, NJ.

Yegnanarayana, B., Murthy, H.A., 1992. Significance of group delay functions in spectrum estimation. IEEE Trans. Signal Process. 40, 2281–2289. doi:10.1109/78.157227

Appendix

Appendix 1 . SSCF Angles comparisons among different Vietnamese V1V2 transitions for each speaker

This part will show the comparison results of three angles (angle12, angle23, angle34) among the different items (/ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/) for each subject with two speech rates (normal and fast rate) (from 3 Vietnamese males (Son, Dat, Khoa) and 3 Vietnamese females (Diep, Yen, Mai)), as follows:

- **SSCF Angle12**

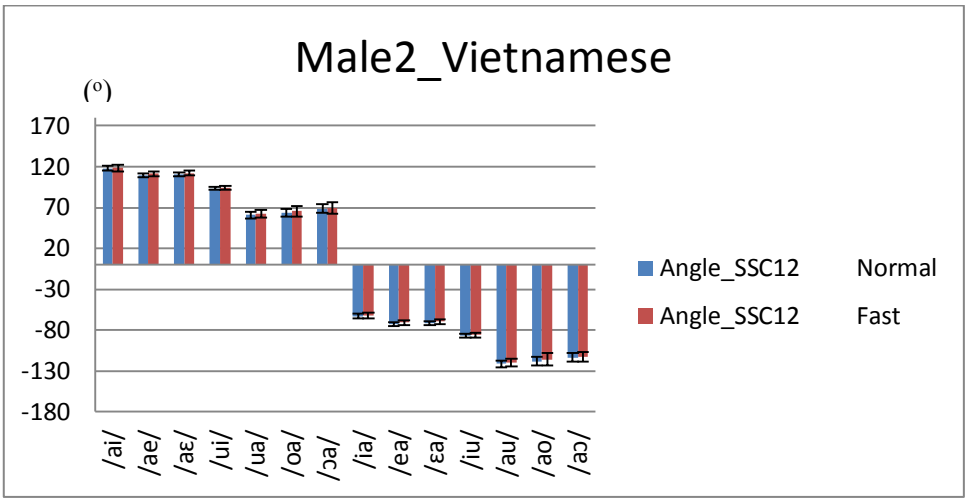


Figure A1-0-1:The average value and standard deviation of angle12 of all items of Vietnamese male2 (Son).

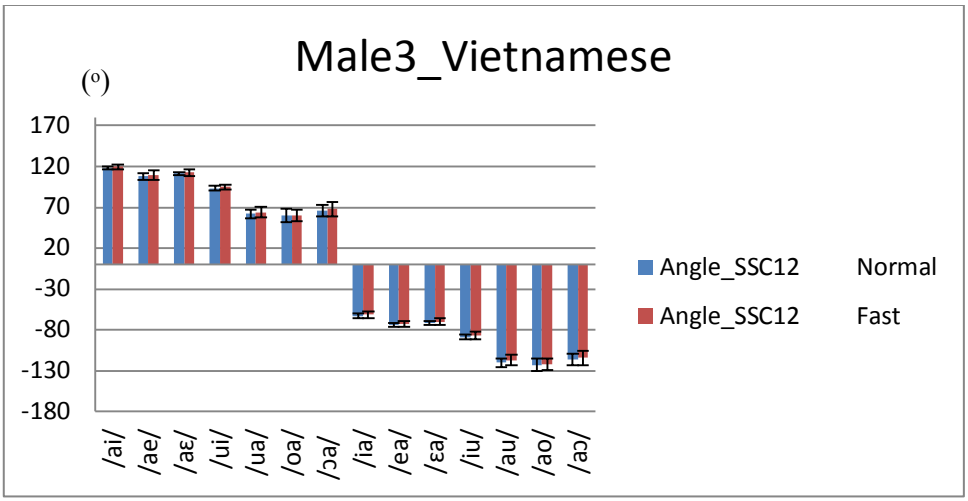


Figure A1-0-2: The average value and standard deviation of angle12 of all items of Vietnamese male3 (Dat).

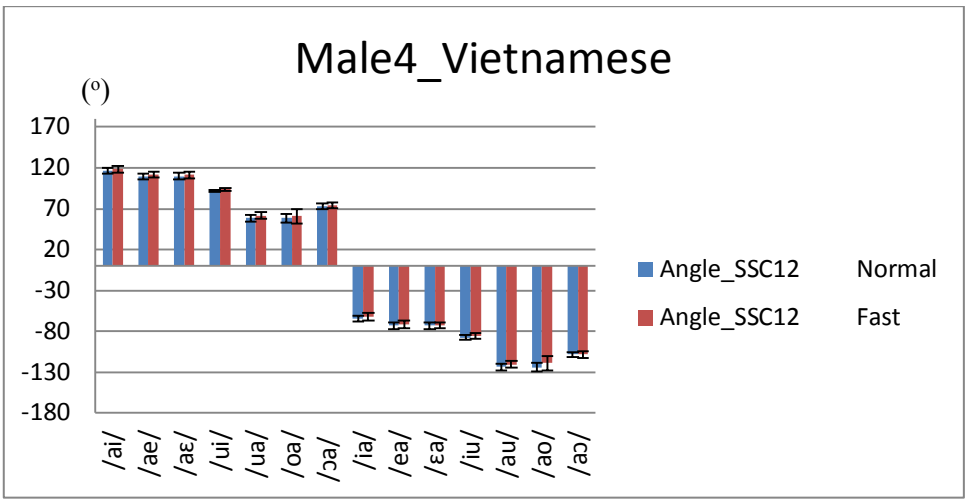


Figure A1-0-3: The average value and standard deviation of angle12 of all items of Vietnamese male 4 (Khoa).

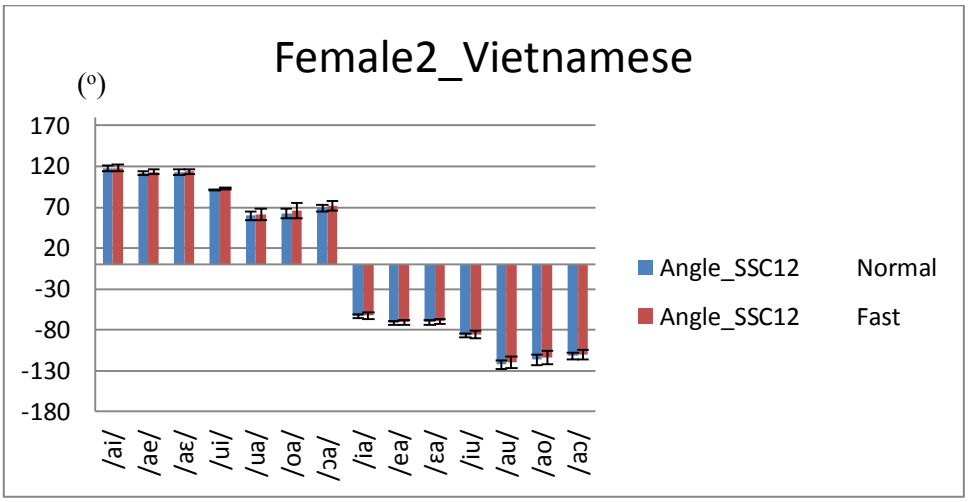


Figure A1-0-4: The average value and standard deviation of angle12 of all items of Vietnamese female 2 (Diep).

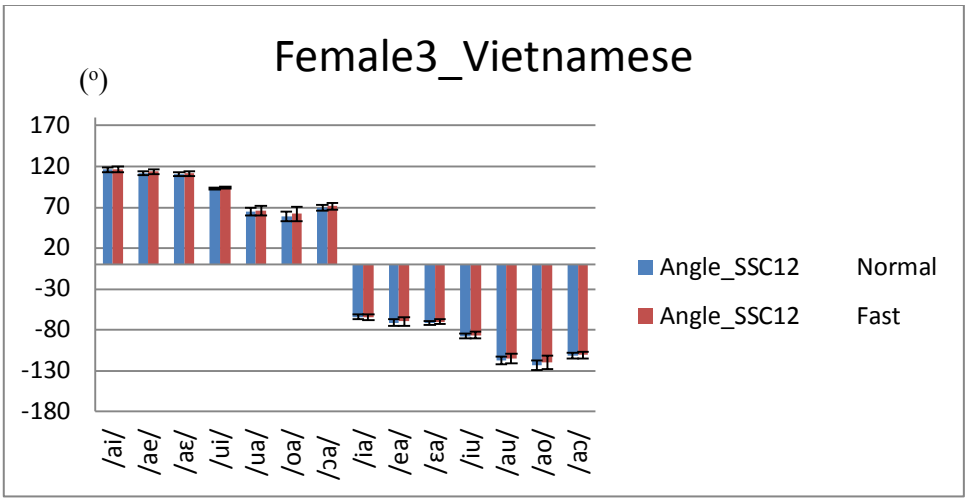


Figure A1-0-5: The average value and standard deviation of angle12 of all items of Vietnamese female 3 (Yen).

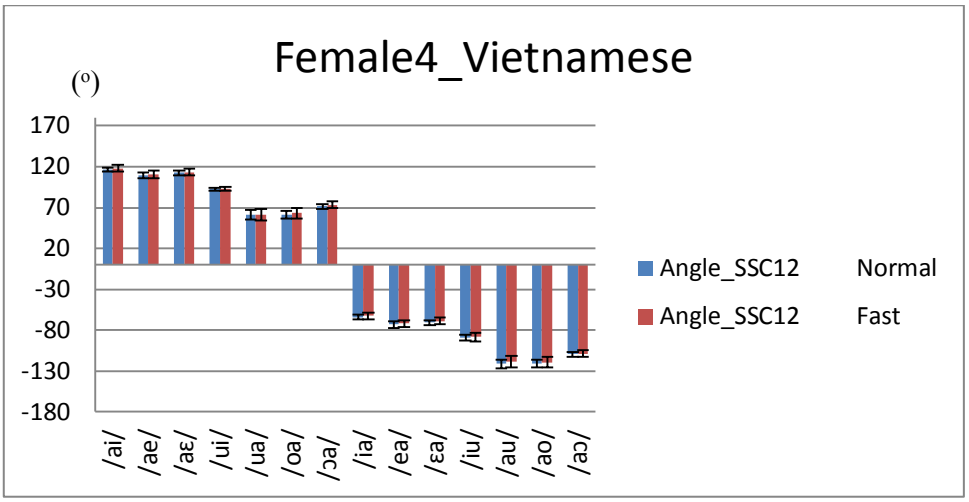


Figure A1-0-6: The average value and standard deviation of angle12 of all items of Vietnamese female 4 (Mai).

❖ **SSCF Angle23**

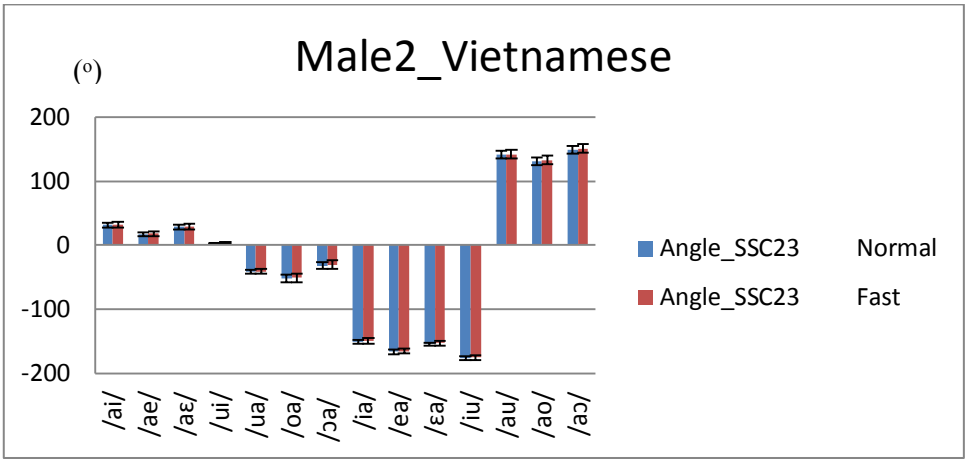


Figure A1-0-7: The average value and standard deviation of angle23 of all items of Vietnamese male 2 (Son).

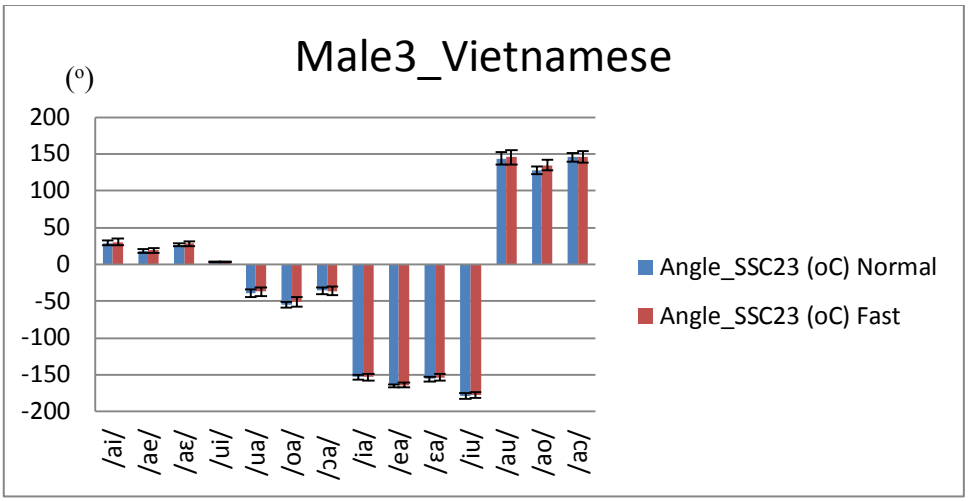


Figure A1-0-8: The average value and standard deviation of angle23 of all items of Vietnamese male 3 (Dat).

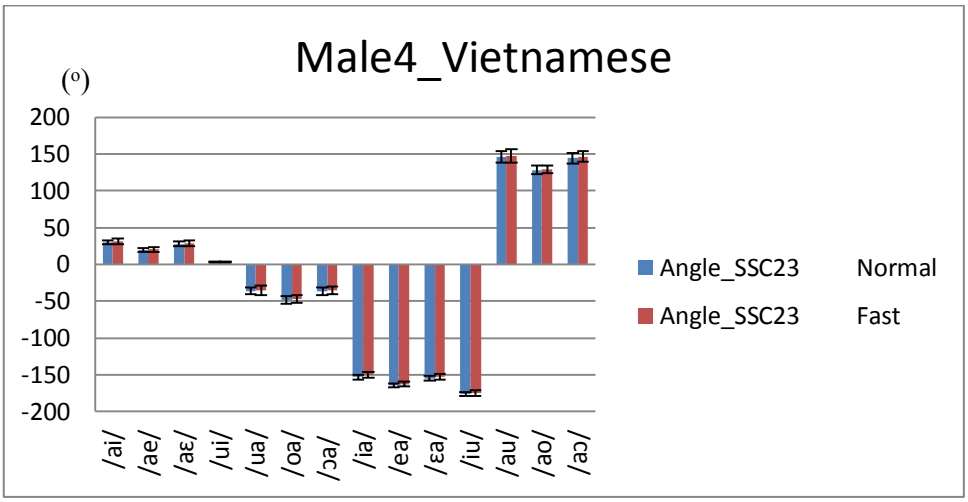


Figure A1-0-9: The average value and standard deviation of angle23 of all items of Vietnamese male 4 (Khoa).

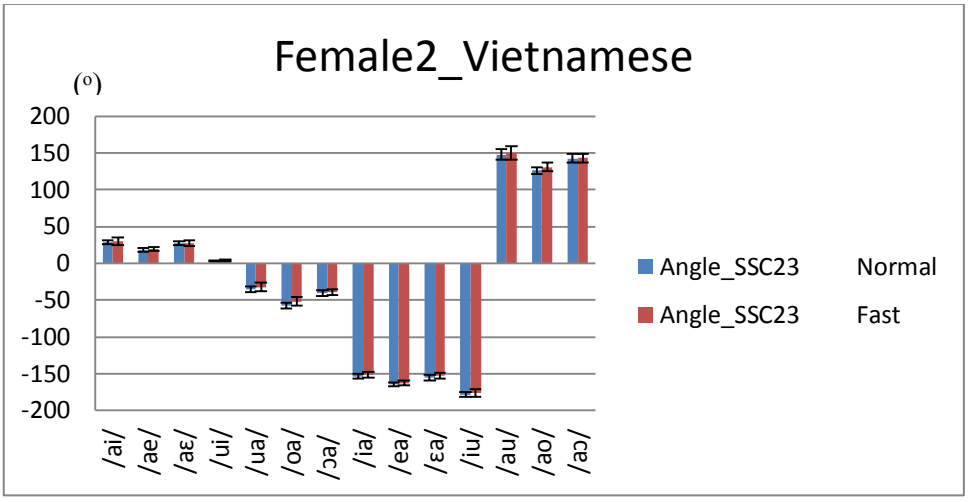


Figure A1-0-10: The average value and standard deviation of angle23 of all items of Vietnamese female 2 (Diep).

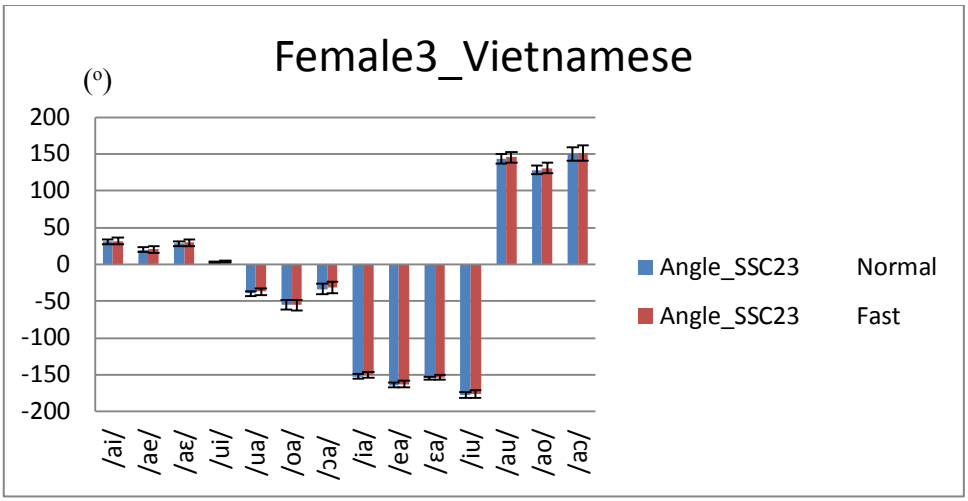


Figure A1-0-11: The average value and standard deviation of angle23 of all items of Vietnamese female 3 (Yen).

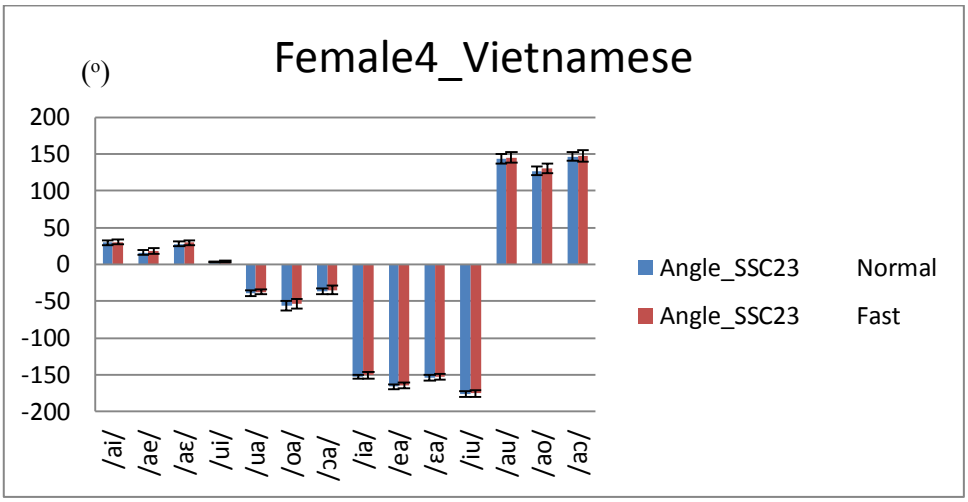


Figure A1-0-12: The average value and standard deviation of angle23 of all items of Vietnamese female 4 (Mai).

❖ **SSCF Angle34**

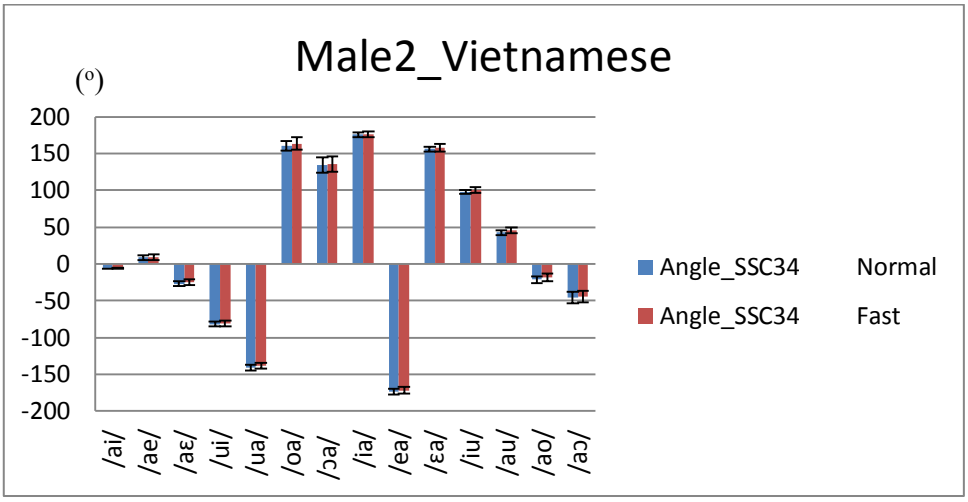


Figure A1-0-13: The average value and standard deviation of angle34 of all items of Vietnamese male 2 (Son).

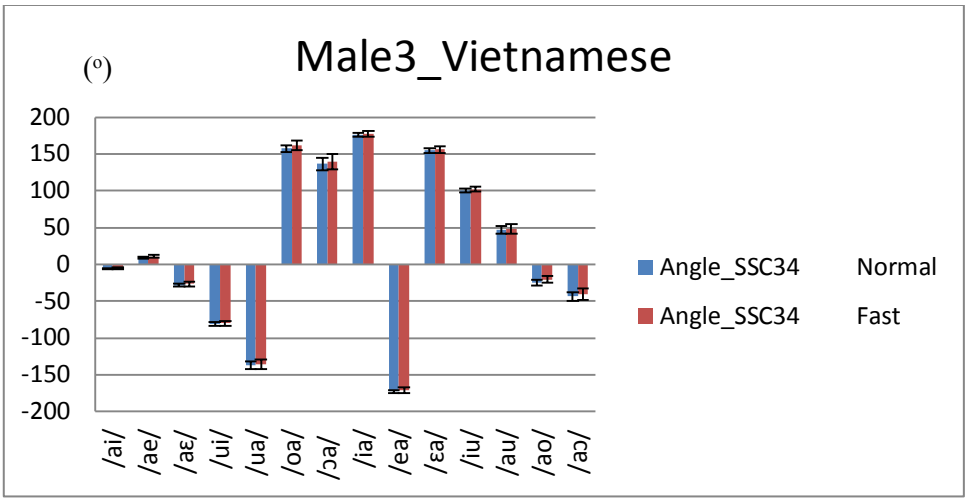


Figure A1-0-14: The average value and standard deviation of angle34 of all items of Vietnamese male 3 (Dat).

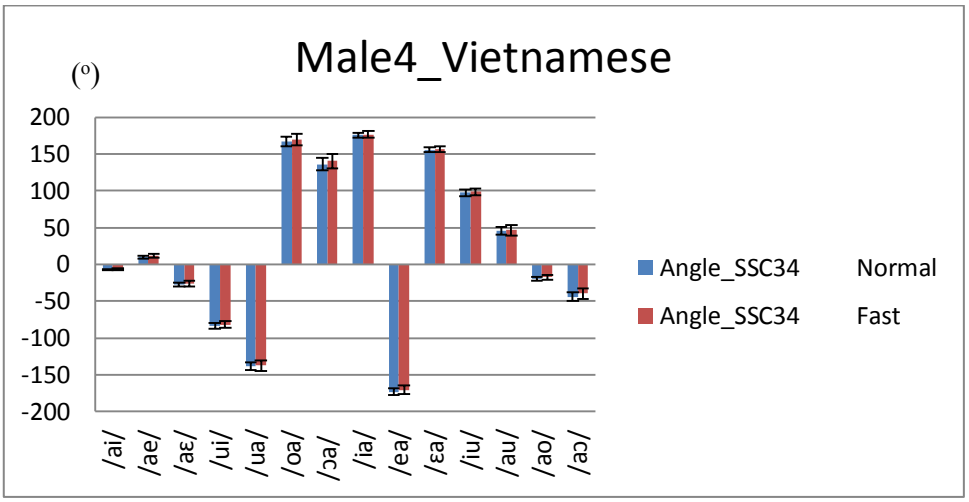


Figure A1-0-15: The average value and standard deviation of angle34 of all items of Vietnamese male 4 (Khoa).

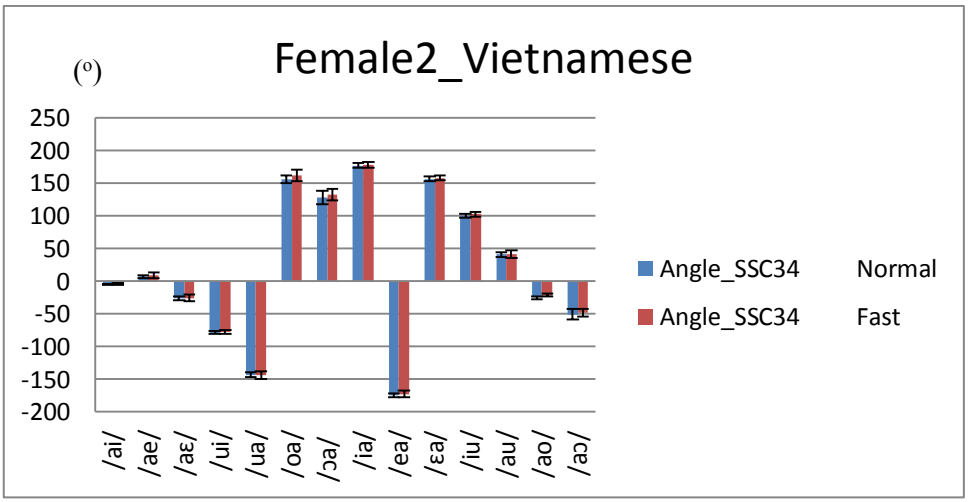


Figure A1-0-16: The average value and standard deviation of angle34 of all items of Vietnamese female 2 (Diep).

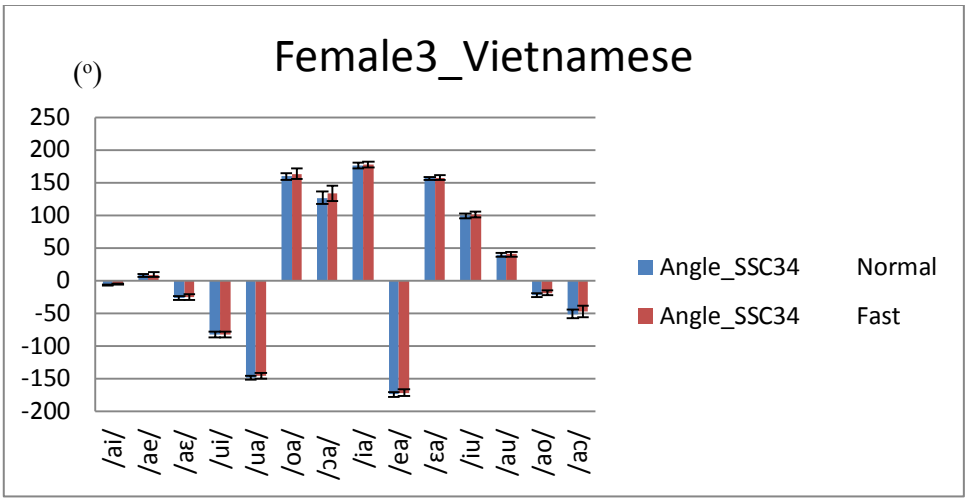


Figure A1-0-17: The average value and standard deviation of angle34 of all items of Vietnamese female 3 (Yen).

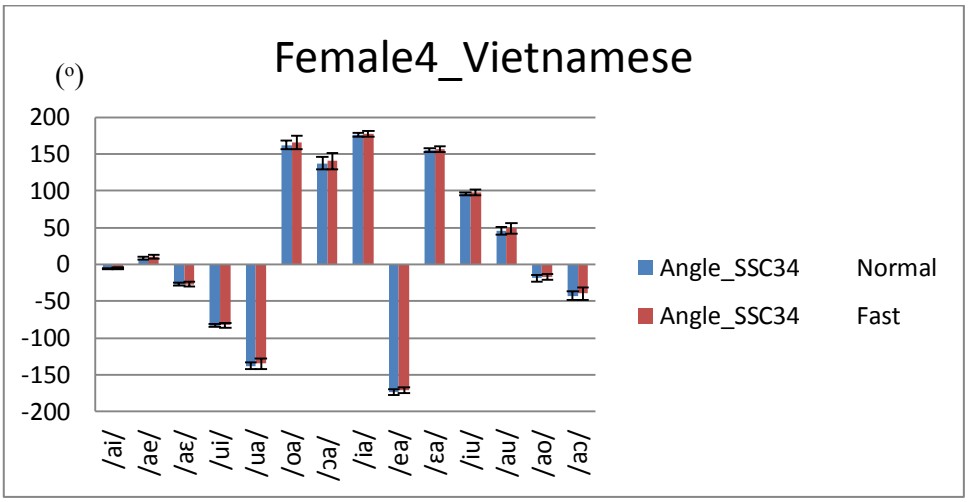


Figure A1-0-18: The average value and standard deviation of angle34 of all items of Vietnamese female 4 (Mai).

Appendix 2 . SSCF Angles comparisons with same Vietnamese V1V2 transitions produced by different speakers

This part will present the comparison of three SSCF Angles (angle12, angle23, angle34) of the same transition produced by eight Vietnamese native speakers (4 males + 4 females) at both normal and fast rate.

A2-1, /aɛ/ sequence

- ***SSCF Angle12 of /aɛ/***

Figure A2-0-1 presents the average results of SSCF Angle12 of the transition /aɛ/ of eight Vietnamese speakers (four males and four females). Some comments are given, as follows:

- For each subject, SSCF Angle12 of transition /aɛ/ is more or less the same with both normal and fast rate with its small standard deviation.
- The SSCF Angle12 of transition /aɛ/ produced by eight speakers are more or less the same value.

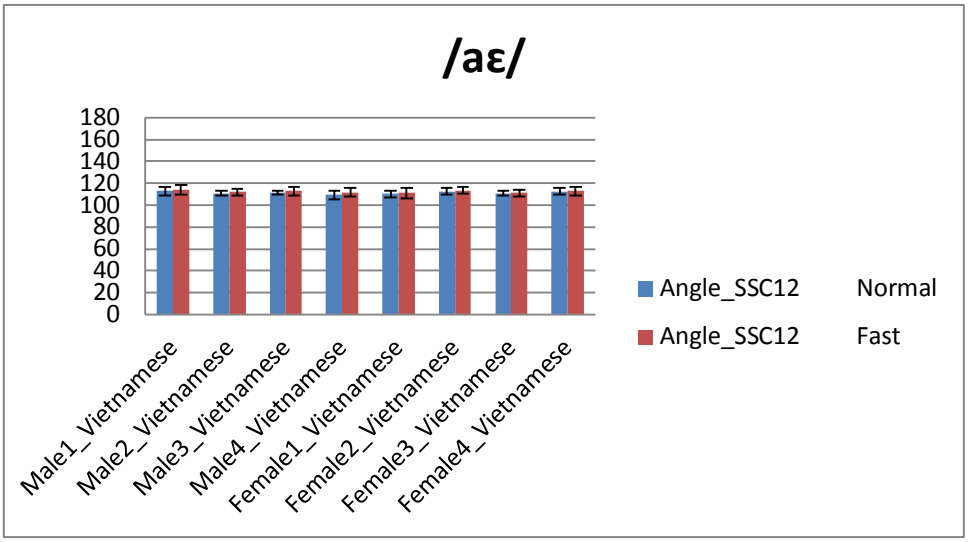


Figure A2-0-1: The average value and standard deviation of angle12 of /aɛ/ with all 8 Vietnamese subjects (4 males + 4 females).

- ***SSCF Angle23 of /aɛ/***

Figure A2-0-2 presents the average results of SSCF Angle23 of the transition /aɛ/ produced by eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, SSCF Angle23 of transition /aɛ/ is more or less the same with both normal and fast rate with its small standard deviation.
- The SSCF Angle23 of transition /aɛ/ among eight speakers are more or less the same value.

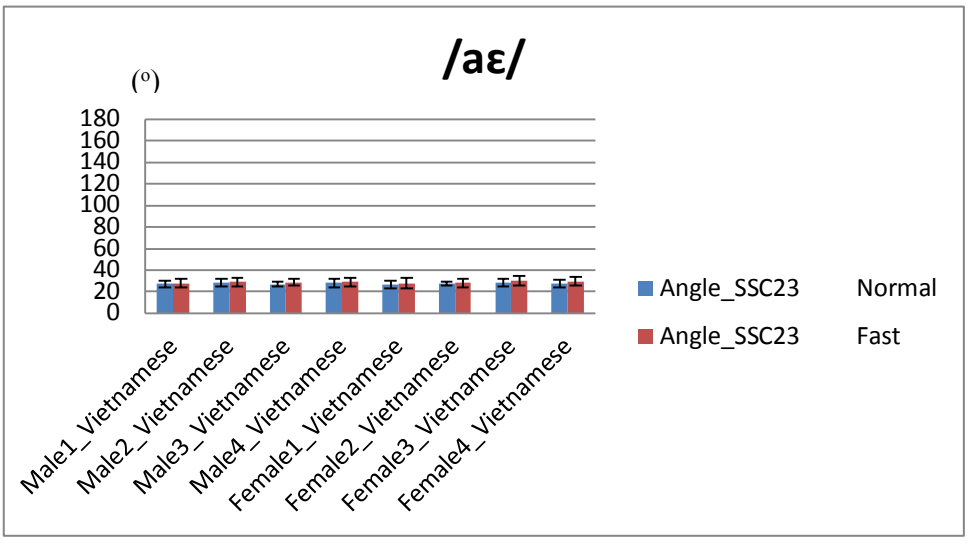


Figure A2-0-2: The average value and standard deviation of angle23 of /aɛ/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle34 of /aɛ/***

Figure A2-0-3 presents the mean results of angle34 of the transition /aɛ/ produced by eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, SSCF Angle34 of transition /aɛ/ is more or less the same with both normal and fast rate.
- The SSCF Angle34 of transition /aɛ/ among eight speakers are more or less the same value.

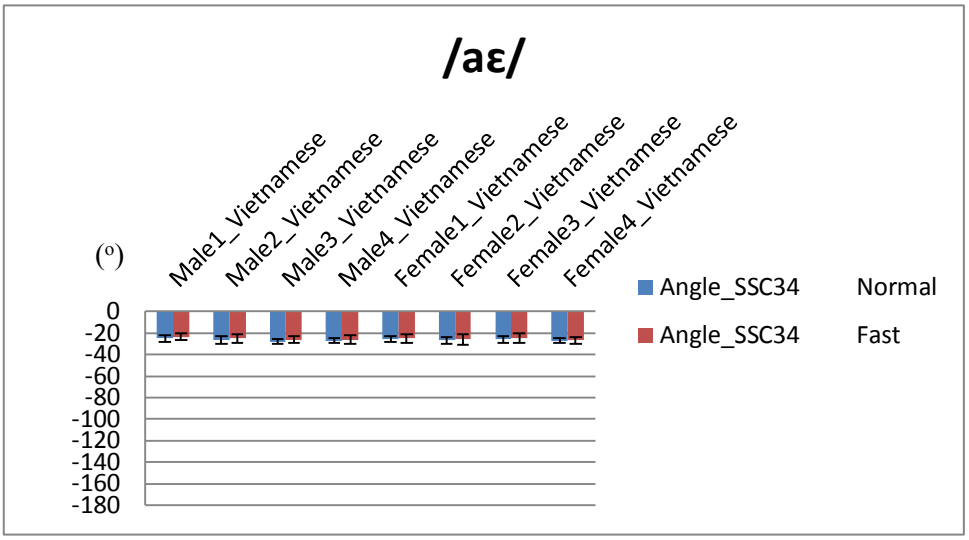


Figure A2-0-3: The average value and standard deviation of angle23 of /aɛ/ with all 8 Vietnamese subjects (4 males + 4 females).

A2-2, /ae/ sequence

❖ ***SSCF Angle12 of /ae/***

Figure A2-0-4 presents the results of angle12 of the transition /ae/ produced by eight Vietnamese speakers (four males and four females). Some comments are given, as follows:

- For each subject, SSCF Angle12 of transition /ae/ is more or less the same with both normal and fast rate with its small standard deviation.
- The SSCF Angle12 of transition /ae/ among eight speakers are more or less the same value.

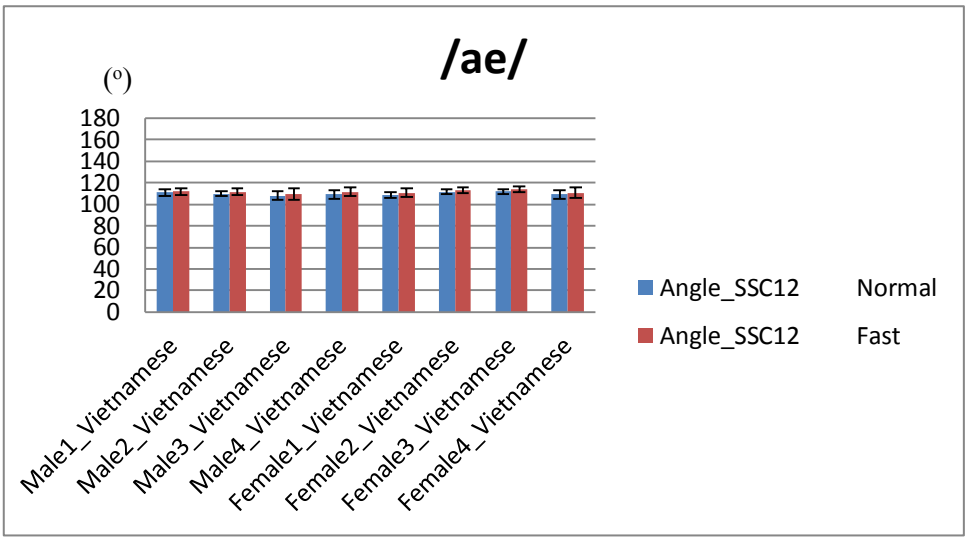


Figure A2-0-4: The average value and standard deviation of angle12 of /ae/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle23 of /ae/***

Figure A2-0-5 presents the mean results of angle23 of the transition /ae/ produced by eight Vietnamese speakers (four males and four females).

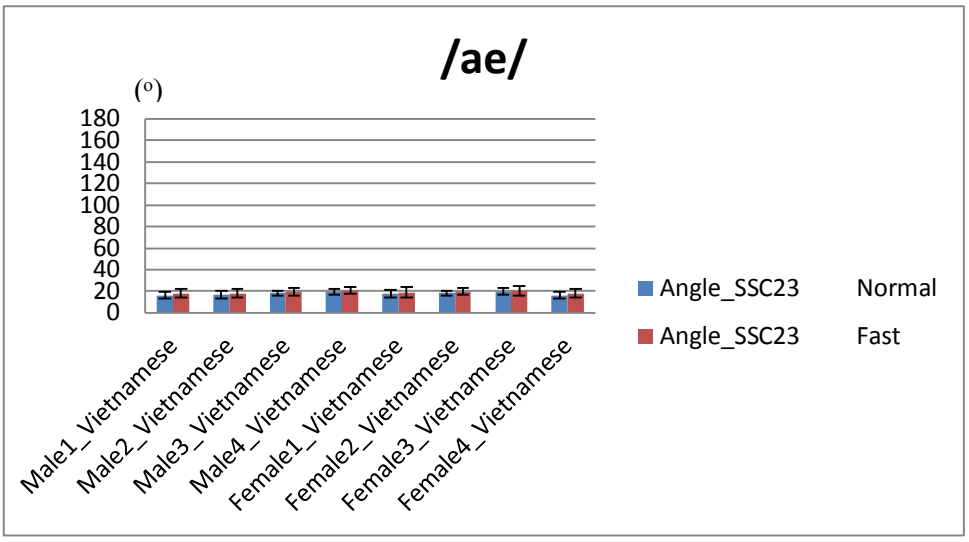


Figure A2-0-5: The average value and standard deviation of angle23 of /ae/ with all 8 Vietnamese subjects (4 males + 4 females).

Some comments are given, as following:

- For each subject, SSCF Angle23 of transition /ae/ is more or less the same with both normal and fast rate.
- The SSCF Angle23 of transition /ae/ among eight speakers are more or less the same value.

❖ ***SCF angle34 of /ae/***

Figure A2-0-6 presents the average results of SSCF Angle34 of the transition /ae/ of eight Vietnamese speakers (four males and four females). Some comments are given as following:

- For each subject, although the standard deviation is a bit large, but the SSCF Angle34 of transition /ae/ is still more or less the same with both normal and fast rate.
- The SSCF Angle34 of transition /ae/ among eight speakers are more or less the same value.

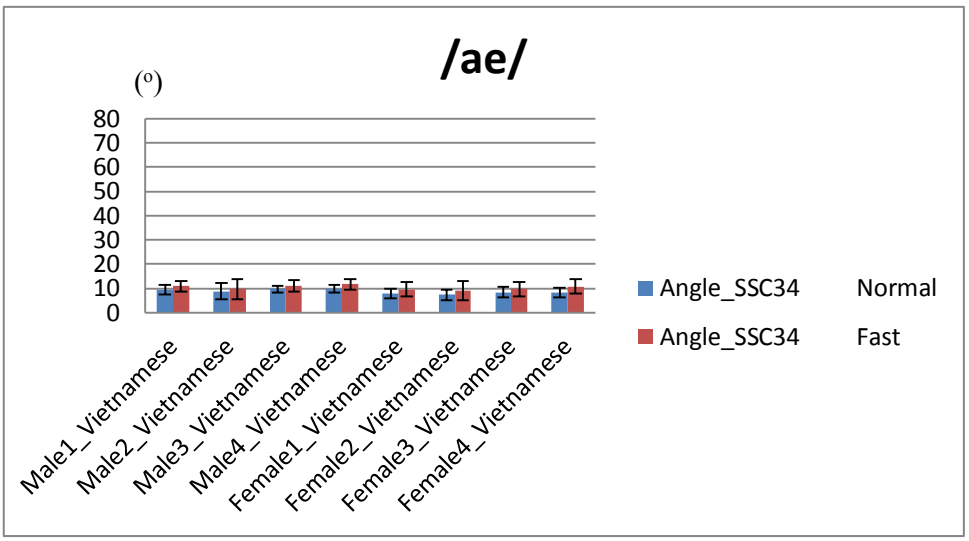


Figure A2-0-6: The average value and standard deviation of angle34 of /ae/ with all 8 Vietnamese subjects (4 males + 4 females).

A2-3, /ua/ sequence

❖ ***SSCF Angle12 of /ua/***

Figure A2-0-7 presents the average results of angle12 of the transition /ua/ of eight Vietnamese speakers (four males and four females). Some comments are given:

- For each subject, angle12 of transition /ua/ is more or less the same with both normal and fast rate;
- The angle12 of transition /ua/ among eight speakers are more or less the same value.

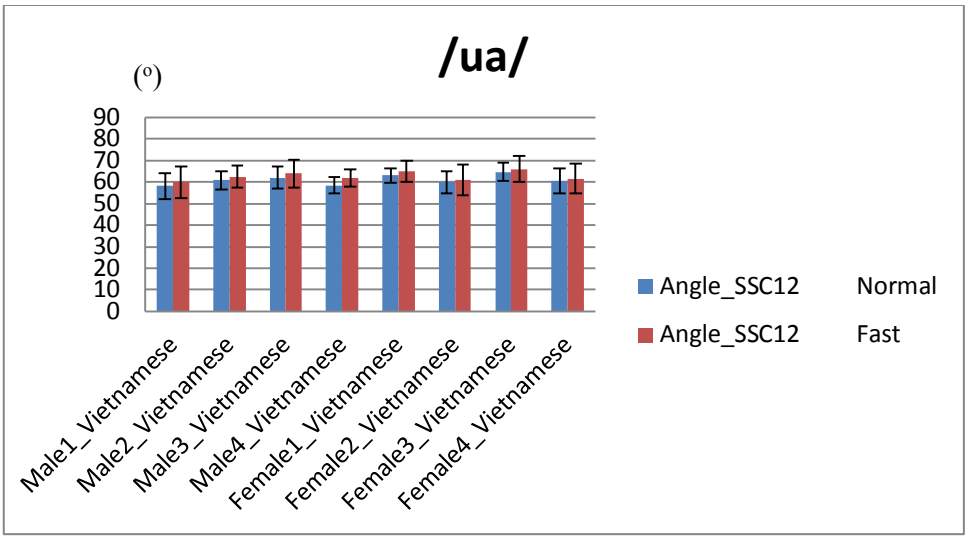


Figure A2-0-7: The average value and standard deviation of angle12 of /ua/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle23 of /ua/***

Figure A2-0-8 presents the average results of angle23 of the transition /ua/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle23 of transition /ua/ is more or less the same with both normal and fast rate.
- The angle23 of transition /ua/ among eight speakers are more or less the same value.

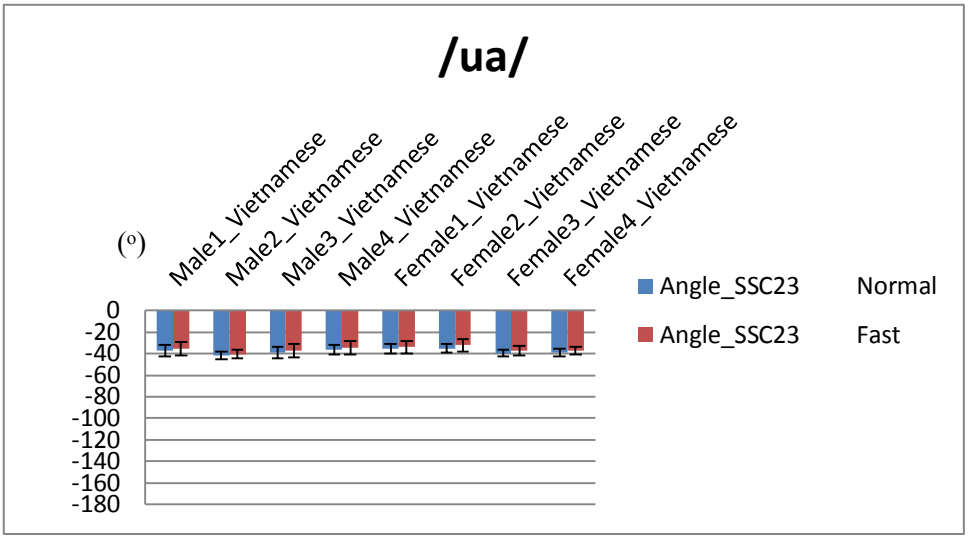


Figure A2-0-8: The average value and standard deviation of angle23 of /ua/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle34 of /ua/***

Figure A2-0-9 presents the average results of angle34 of the transition /ua/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle34 of transition /ua/ is more or less the same with both normal and fast rate with its small standard deviation.
- The angle34 of transition /ua/ among eight speakers are more or less the same value.

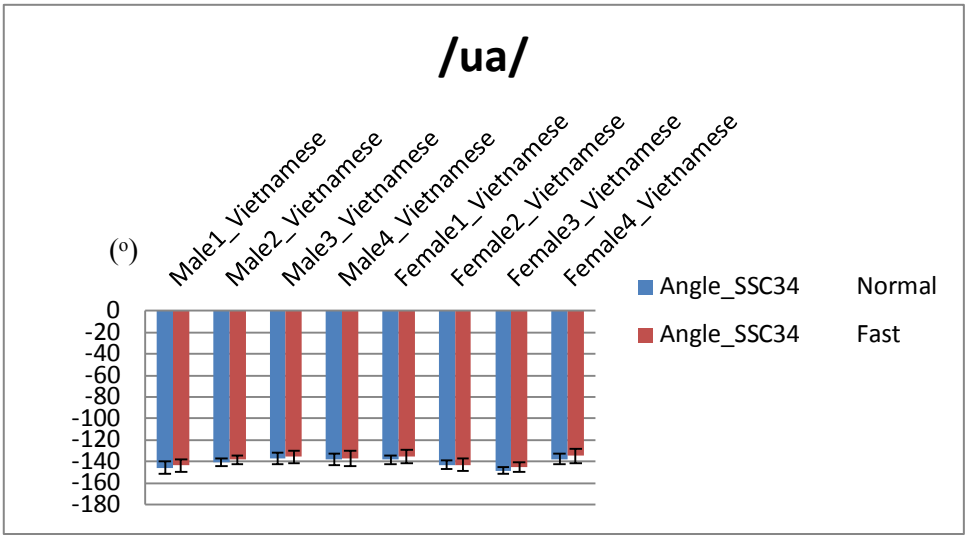


Figure A2-0-9: The average value and standard deviation of angle34 of /ua/ with all 8 Vietnamese subjects (4 males + 4 females).

A2-4, /ui/ seqence

❖ ***SSCF Angle12 of /ui/***

Figure A2-0-10 presents the average results of angle12 of the transition /ui/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle12 of transition /ui/ is more or less the same with both normal and fast rate with its very small standard deviation.
- The angle12 of transition /ui/ among eight speakers are more or less the same value.

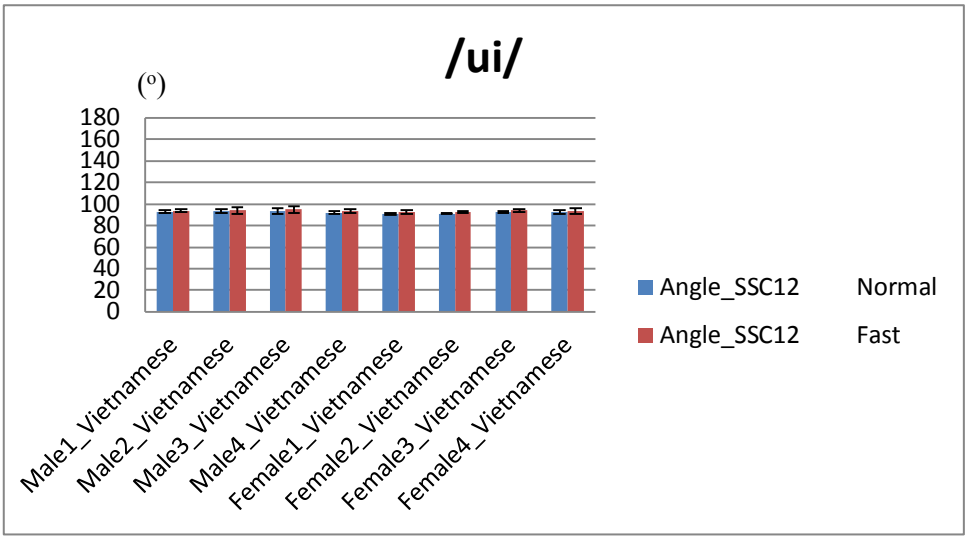


Figure A2-0-10: The average value and standard deviation of angle12 of /ui/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle23 of /ui/***

Figure A2-0-11 presents the average results of angle23 of the transition /ui/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, although the standard deviation is large, but basically, the angle23 of transition /ui/ is still more or less the same with both normal and fast rate.
- The angle23 of transition /ui/ among eight speakers are more or less the same value.

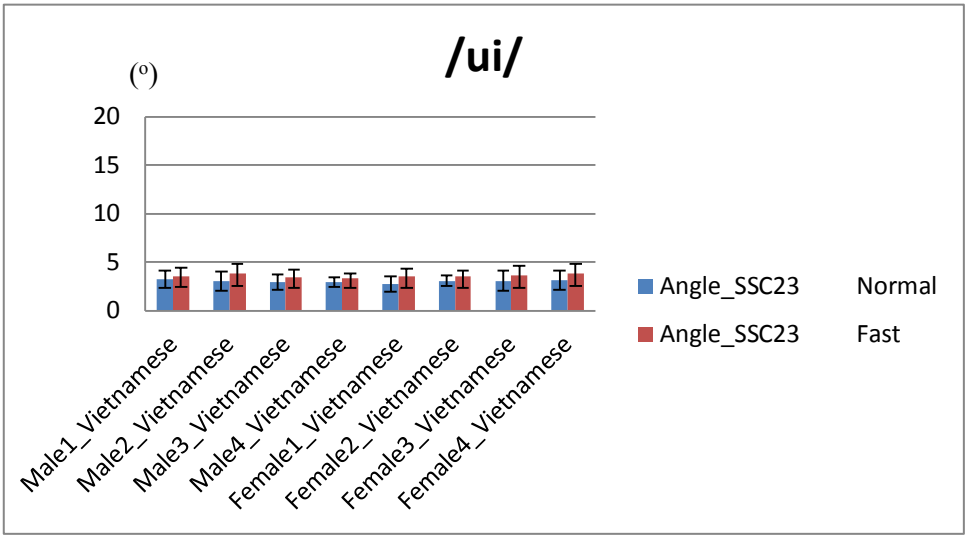


Figure A2-0-11: The average value and standard deviation of angle23 of /ui/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle34 of /ui/***

Figure A2-0-12 presents the average results of angle34 of the transition /ui/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle34 of transition /ui/ is more or less the same with both normal and fast rate with its standard deviation.
- The angle34 of transition /ui/ among eight speakers are more or less the same value.

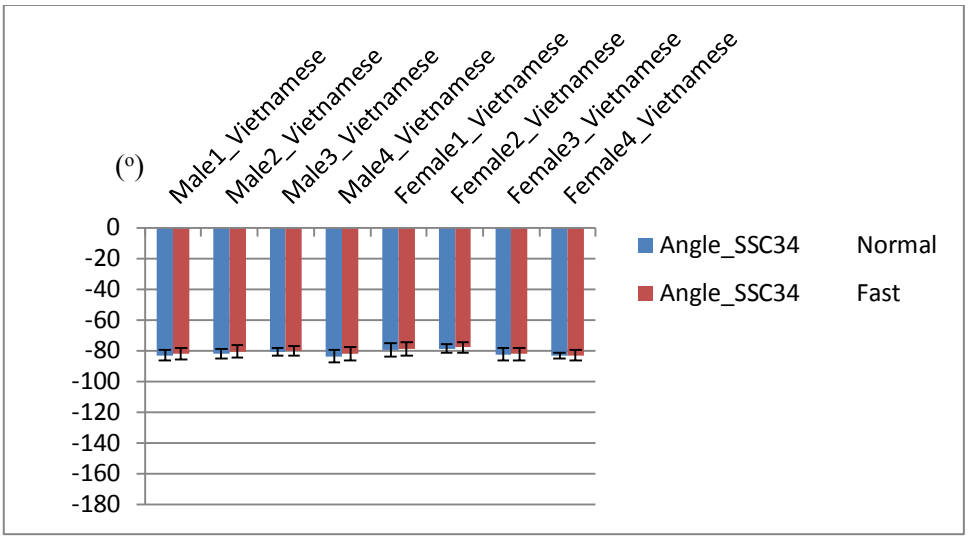


Figure A2-0-12: The average value and standard deviation of angle34 of /ui/ with all 8 Vietnamese subjects (4 males + 4 females).

A2-5, /oa/ sequence

- ***SSCF Angle12 of /oa/***

Figure A2-0-13 presents the average results of angle12 of the transition /oa/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle12 of transition /oa/ is more or less the same with both normal and fast rate.
- The angle12 of transition /oa/ among eight speakers are more or less the same value.

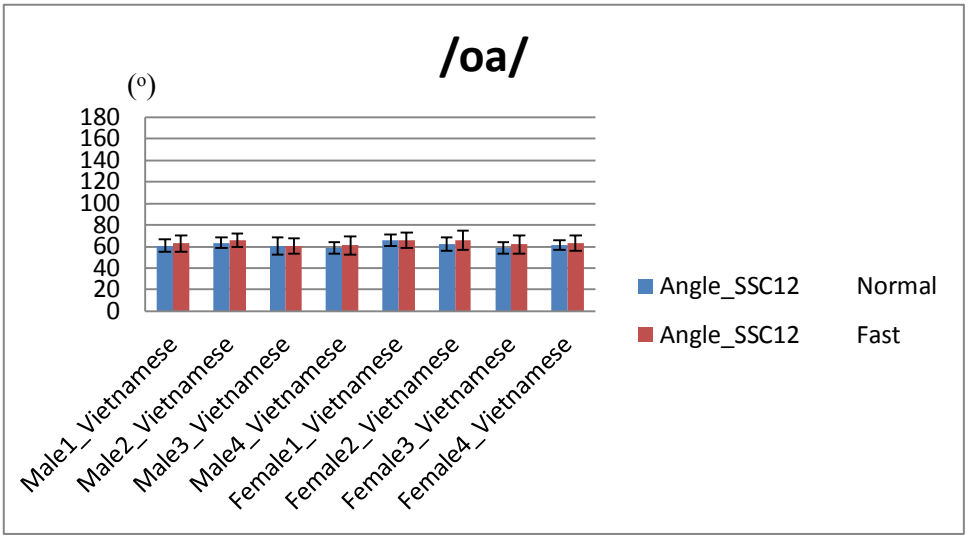


Figure A2-0-13: The average value and standard deviation of angle12 of /oa/ with all 8 Vietnamese subjects (4 males + 4 females).

- ***SSCF Angle23 of /oa/***

Figure A2-0-14 presents the average results of angle23 of the transition /oa/ of eight Vietnamese speakers (four males and four females).

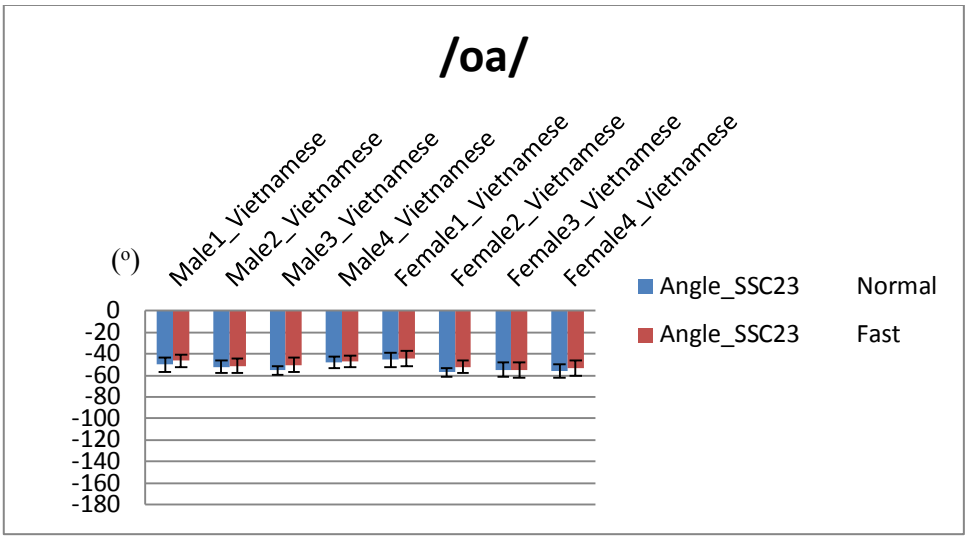


Figure A2-0-14: The average value and standard deviation of angle23 of /oa/ with all 8 Vietnamese subjects (4 males + 4 females).

Some comments are given, as following:

- For each subject, angle23 of transition /oa/ is more or less the same with both normal and fast rate.
- The angle23 of transition /oa/ among eight speakers are more or less the same value.

❖ ***SSCF Angle34 of /oa/***

Figure A2-0-15 presents the average results of angle34 of the transition /oa/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle34 of transition /oa/ is more or less the same with both normal and fast rate.
- The angle34 of transition /oa/ among eight speakers are more or less the same value.

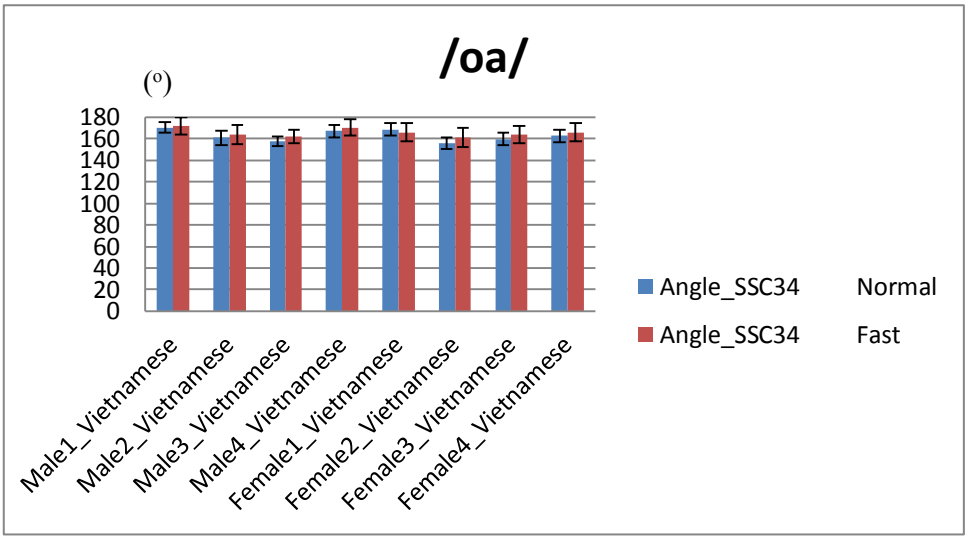


Figure A2-0-15: The average value and standard deviation of angle23 of /oa/ with all 8 Vietnamese subjects (4 males + 4 females).

A2-6, /ɔa/ sequence

❖ ***SSCF Angle12 of /ɔa/***

Figure A2-0-16 presents the average results of angle12 of the transition /ɔa/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle12 of transition /ɔa/ is more or less the same with both normal and fast rate.
- The angle12 of transition /ɔa/ among eight speakers are more or less the same value.

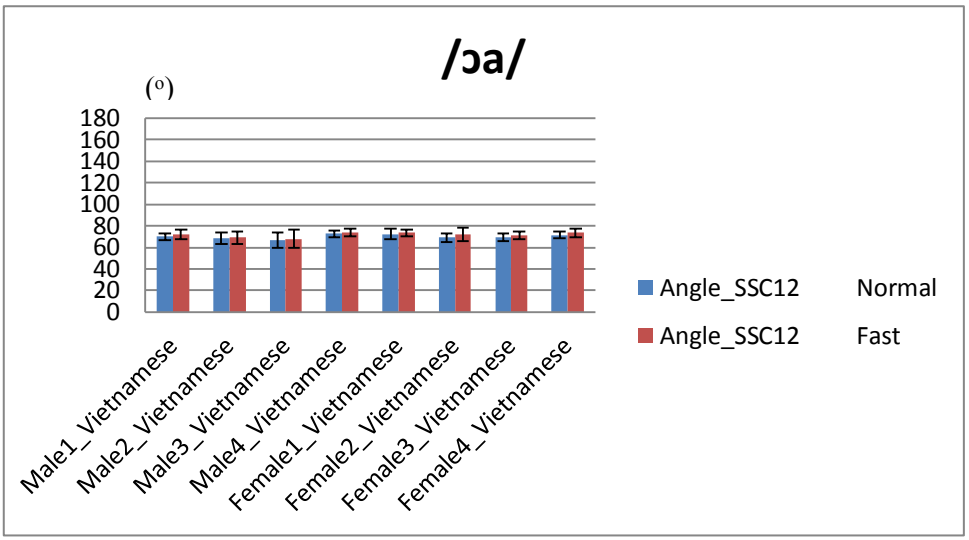


Figure A2-0-16: The average value and standard deviation of angle12 of /ɔa/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle23 of /ɔa/***

Figure A2-0-17 presents the average results of angle23 of the transition /ɔa/ of eight Vietnamese speakers (four males and four females).

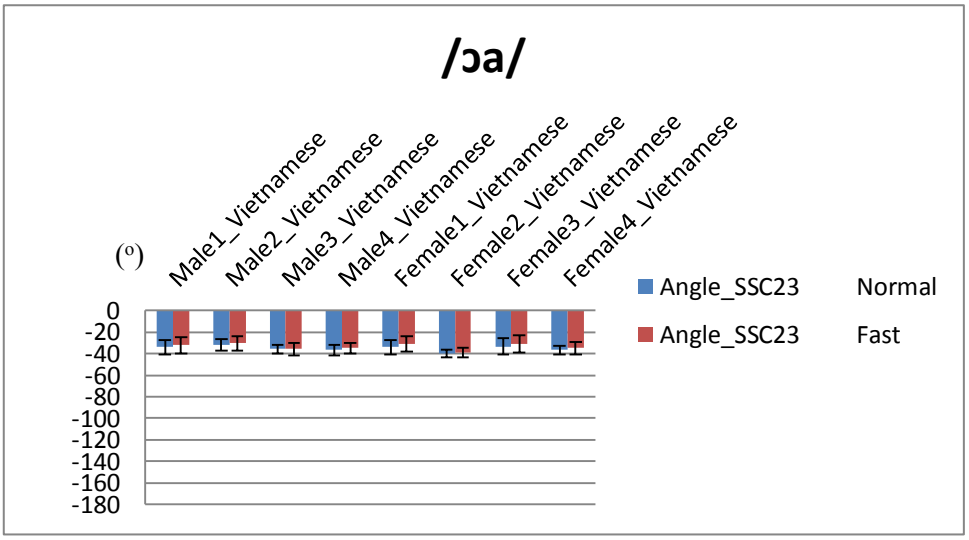


Figure A2-0-17: The average value and standard deviation of angle23 of /ɔa/ with all 8 Vietnamese subjects (4 males + 4 females).

Some comments are given, as following:

- For each subject, angle23 of transition /ɔa/ is more or less the same with both normal and fast rate.
- The angle23 of transition /ɔa/ among eight speakers are more or less the same value.

❖ ***SSCF Angle34 of /ɔa/***

Figure A2-0-18 presents the average results of angle34 of the transition /ɔa/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle34 of transition /ɔa/ is more or less the same with both normal and fast rate.
- The angle34 of transition /ɔa/ among eight speakers are more or less the same value.

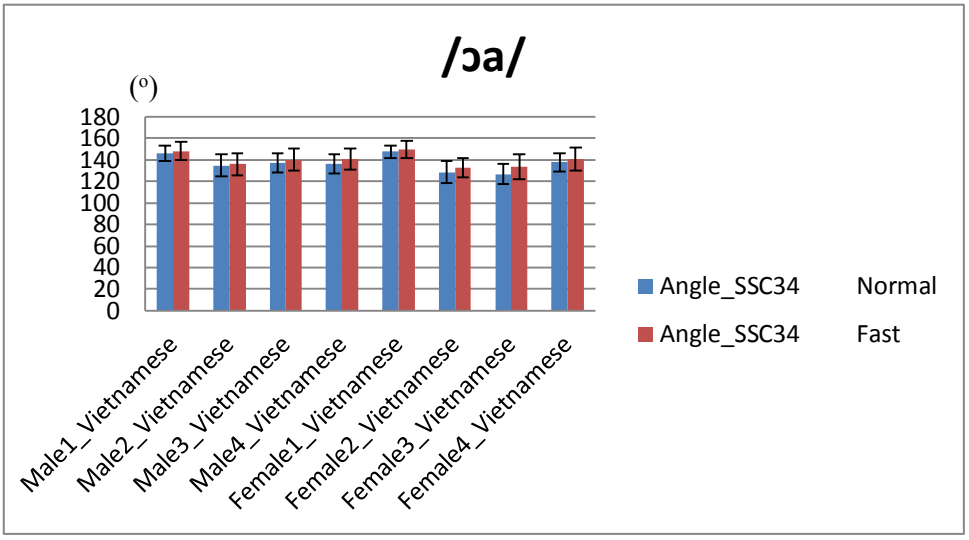


Figure A2-0-18: The average value and standard deviation of angle34 of /ɔa/ with all 8 Vietnamese subjects (4 males + 4 females).

A2-7, /ia/ sequence

❖ ***SSCF Angle12 of /ia/***

Figure A2-0-19 presents the average results of angle12 of the transition /ia/ of eight Vietnamese speakers (four males and four females). Some comments are given, as follows:

- For each subject, angle12 of transition /ia/ is more or less the same with both normal and fast rate.
- The angle12 of transition /ia/ among eight speakers are more or less the same value.

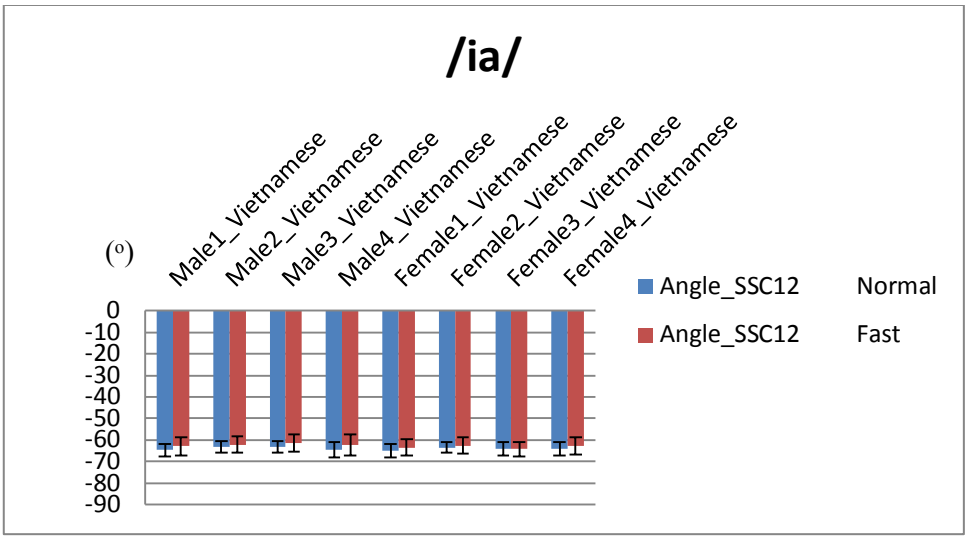


Figure A2-0-19: The average value and standard deviation of angle12 of /ia/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle23 of /ia/***

Figure A2-0-20 presents the average results of angle23 of the transition /ia/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle23 of transition /ia/ is more or less the same with both normal and fast rate with its small standard deviation.
- The angle23 of transition /ia/ among eight speakers are more or less the same value.

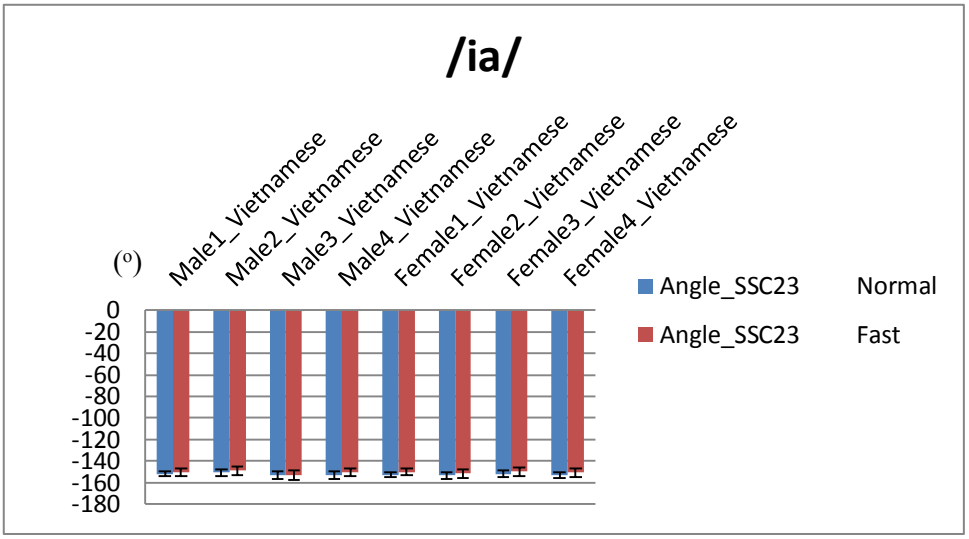


Figure A2-0-20: The average value and standard deviation of angle23 of /ia/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle34 of /ia/***

Figure A2-0-21 presents the average results of angle34 of the transition /ia/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle34 of transition /ia/ is more or less the same with both normal and fast rate with its very small standard deviation.
- The angle34 of transition /ia/ among eight speakers are more or less the same value.

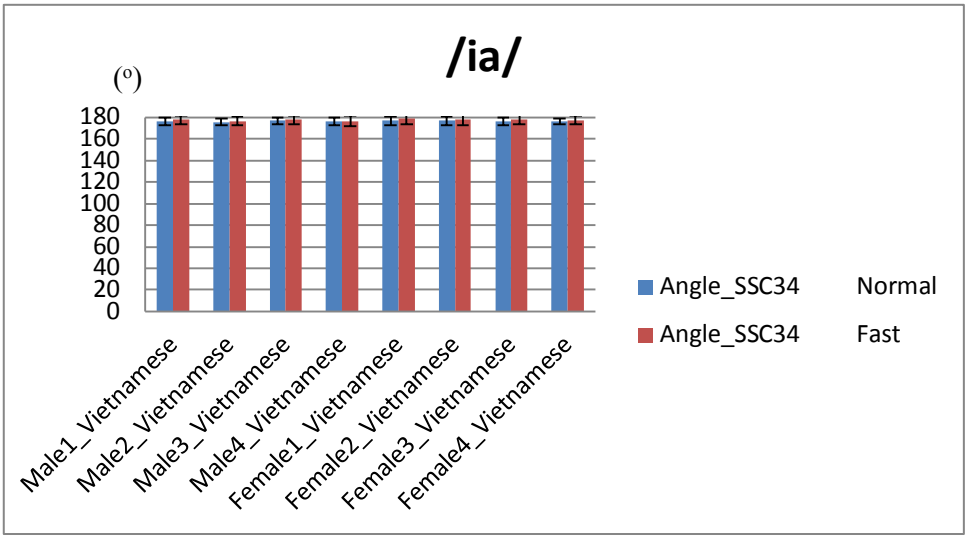


Figure A2-0-21: The average value and standard deviation of angle34 of /ia/ with all 8 Vietnamese subjects (4 males + 4 females).

A2-8, /ɛa/ sequence

❖ ***SSCF Angle12 of /ɛa/***

Figure A2-0-22 presents the mean results of angle12 of the transition /ɛa/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle12 of transition /ɛa/ is more or less the same with both normal and fast rate.
- The angle12 of transition /ɛa/ among eight speakers are more or less the same value.

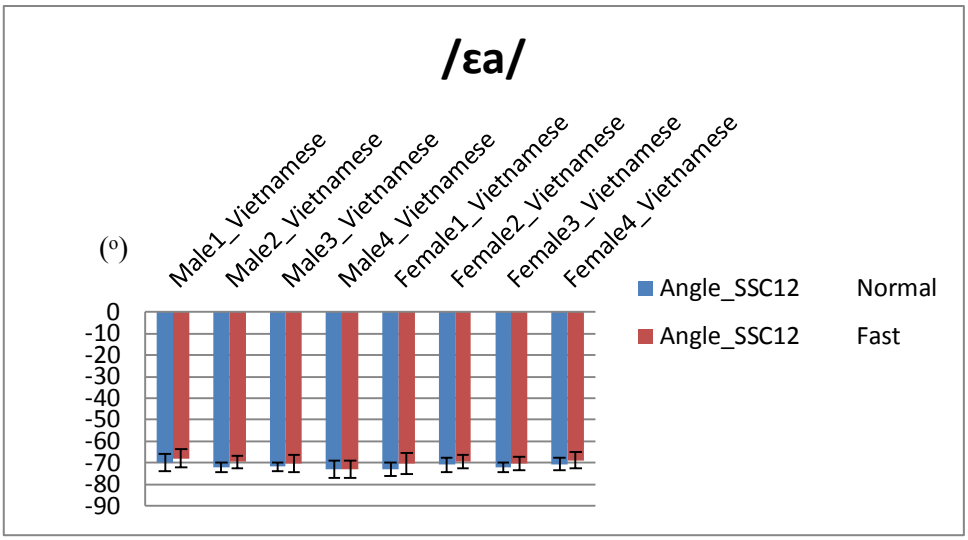


Figure A2-0-22: The average value and standard deviation of angle12of /ɛa/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle23 of /ɛa/***

Figure A2-0-23 presents the average results of angle23 of the transition /ɛa/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle23 of transition /ɛa/ is more or less the same with both normal and fast rate with its very small standard deviation.
- The angle23 of transition /ɛa/ among eight speakers are more or less the same value.

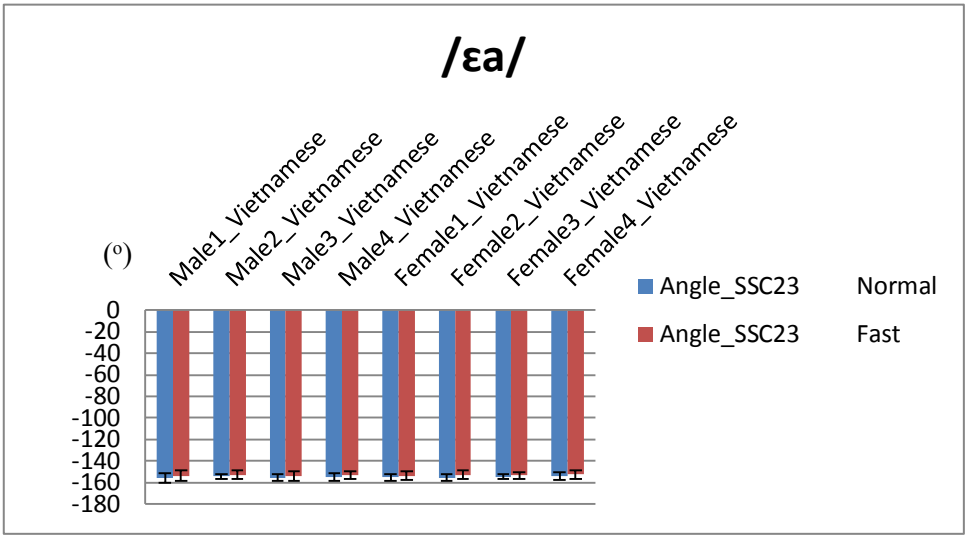


Figure A2-0-23: The average value and standard deviation of angle23 of /ɛa/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle34 of /ɛa/***

Figure A2-0-24 presents the average results of angle34 of the transition /ɛa/ of eight Vietnamese speakers (four males and four females).

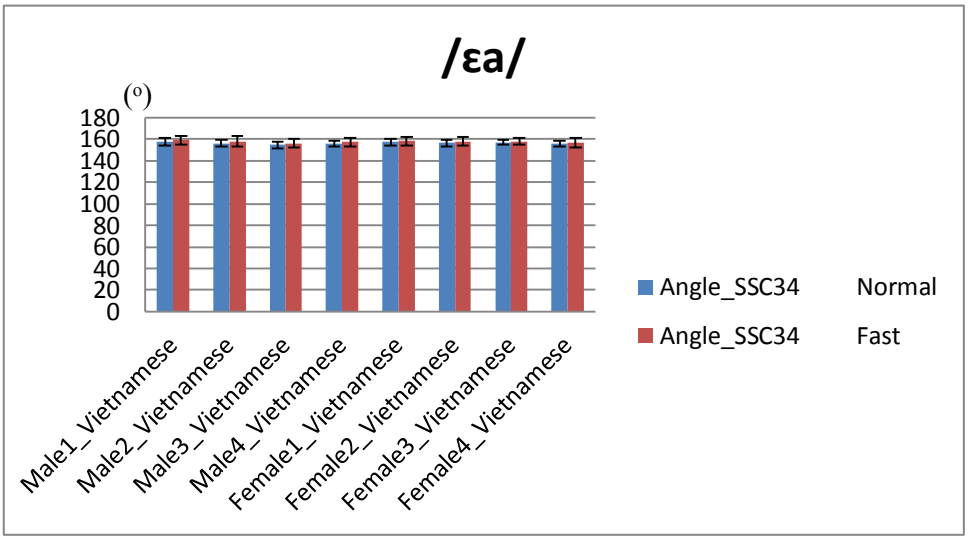


Figure A2-0-24: The average value and standard deviation of angle34 - /ɛa/ with all 8 Vietnamese subjects (4 males + 4 females).

Some comments are given, as following:

- For each subject, angle34 of transition /ɛa/ is more or less the same with both normal and fast rate with its small standard deviation.
- The angle34 of transition /ɛa/ among eight speakers are more or less the same value.

A2-9, /ea/ sequence

❖ ***SSCF Angle12 of /ea/***

Figure A2-0-25 presents the mean results of angle12 of the transition /ea/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle12 of transition /ea/ is more or less the same with both normal and fast rate.
- The angle12 of transition /ea/ among eight speakers are more or less the same value.

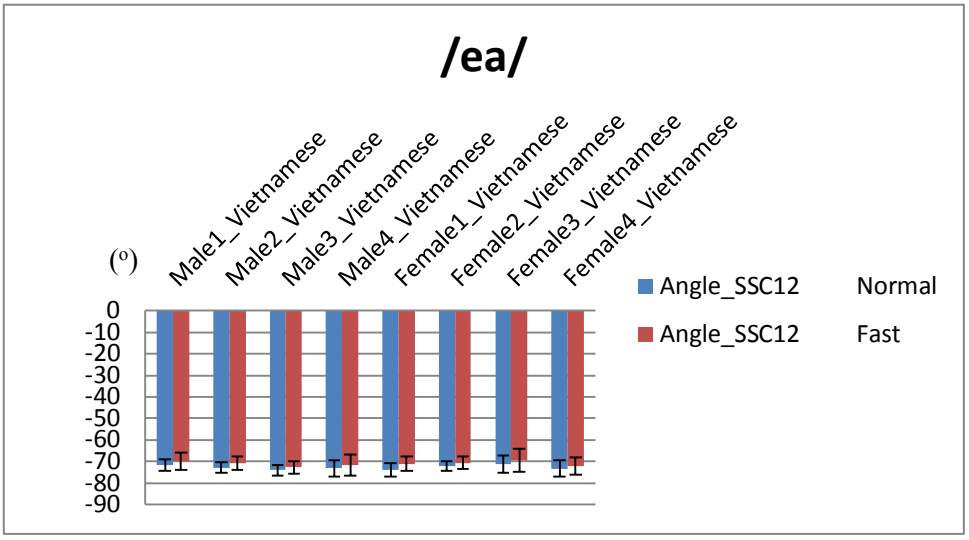


Figure A2-0-25: The average value and standard deviation of angle12 of /ea/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle23 of /ea/***

Figure A2-0-26 presents the average results of angle23 of the transition /ea/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle23 of transition /ea/ is more or less the same with both normal and fast rate with its small standard deviation.
- The angle23 of transition /ea/ among eight speakers are more or less the same value.

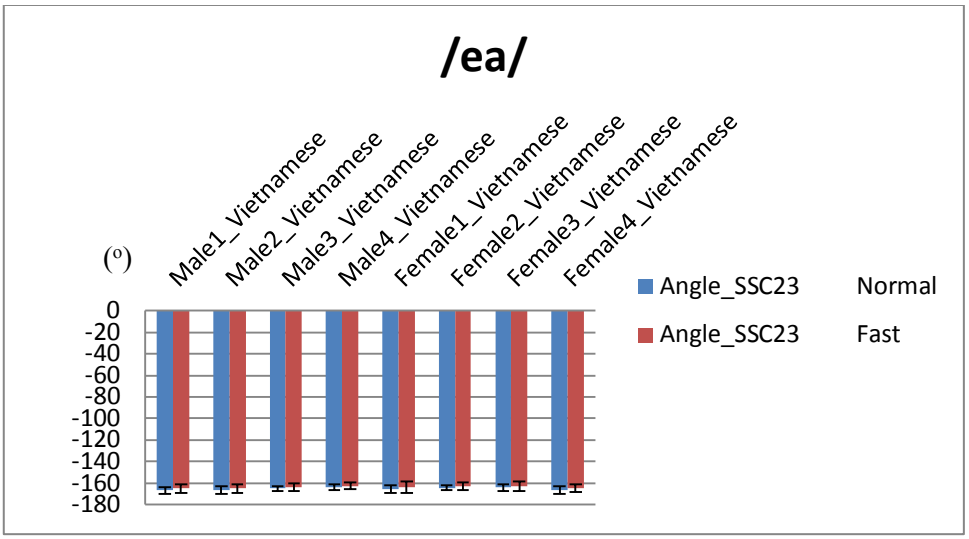


Figure A2-0-26: The average value and standard deviation of angle23 of /ea/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle34 of /ea/***

Figure A2-0-27 presents the mean results of angle34 of the transition /ea/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle34 of transition /ea/ is more or less the same with both normal and fast rate with its small standard deviation.
- The angle34 of transition /ea/ among eight speakers are more or less the same value.

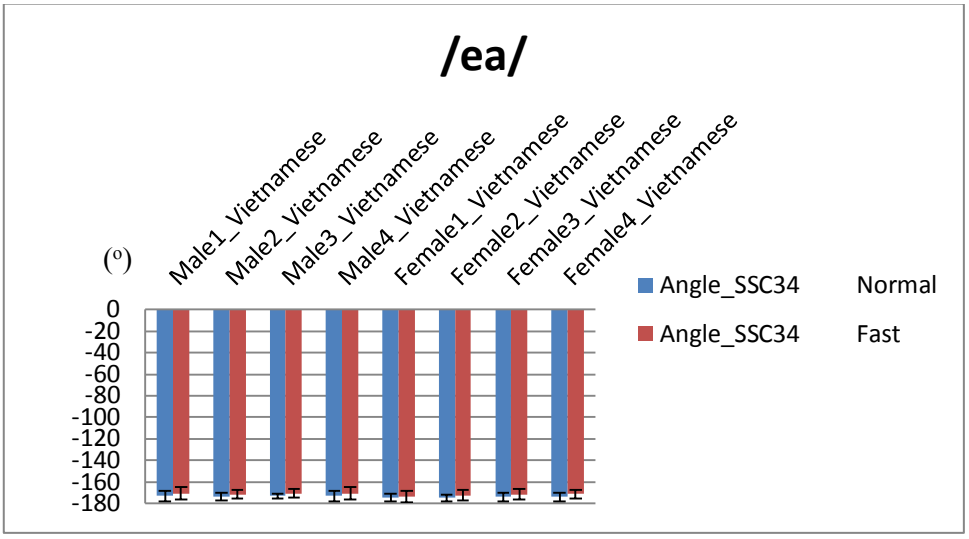


Figure A2-0-27: The average value and standard deviation of angle34 of /ea/ with all 8 Vietnamese subjects (4 males + 4 females).

A2-10, /ao/ sequence

❖ ***SSCF Angle12 of /ao/***

Figure A2-0-28 presents the average results of angle12 of the transition /ao/ of eight Vietnamese speakers (four males and four females).

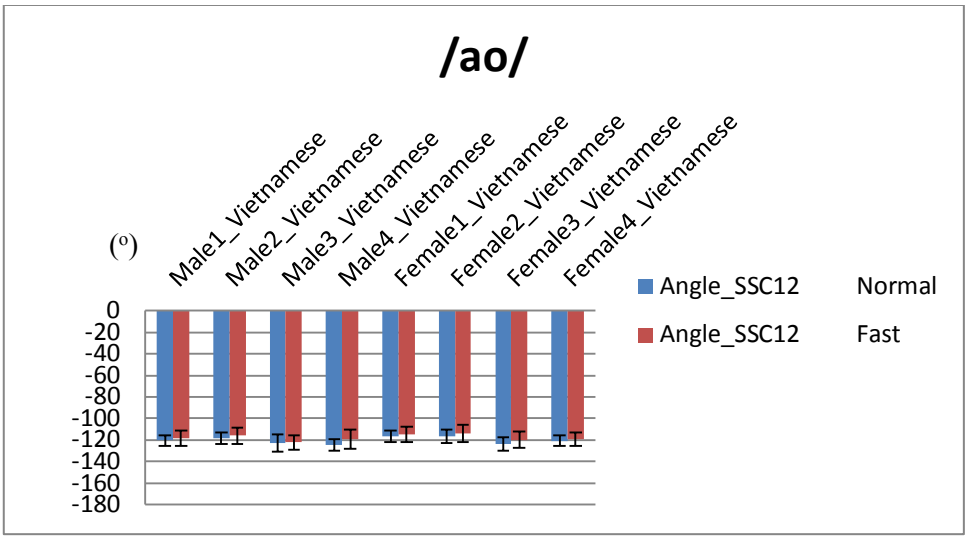


Figure A2-0-28: The average value and standard deviation of angle12 of /ao/ with all 8 Vietnamese subjects (4 males + 4 females).

Some comments are given, as following:

- For each subject, angle12 of transition /ao/ is more or less the same with both normal and fast rate.
- The angle12 of transition /ao/ among eight speakers are more or less the same value.

❖ ***SSCF Angle23 of /ao/***

Figure A2-0-29 presents the average results of angle23 of the transition /ao/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle23 of transition /ao/ is more or less the same with both normal and fast rate.
- The angle23 of transition /ao/ among eight speakers are more or less the same value.

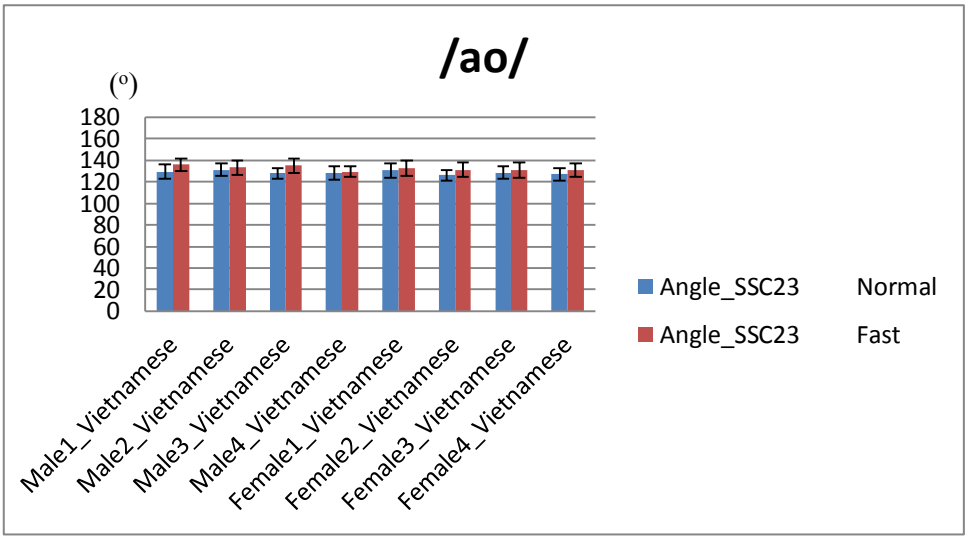


Figure A2-0-29: The average value and standard deviation of angle23 of /ao/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle34 of /ao/***

Figure A2-0-30 presents the average results of angle34 of the transition /ao/ of eight Vietnamese speakers (four males and four females). Some comments are given: (i) for each subject, angle34 of transition /ao/ is more or less the same with both normal and fast rate; (ii) the angle34 of transition /ao/ among eight speakers are not same value, but these differences here are small.

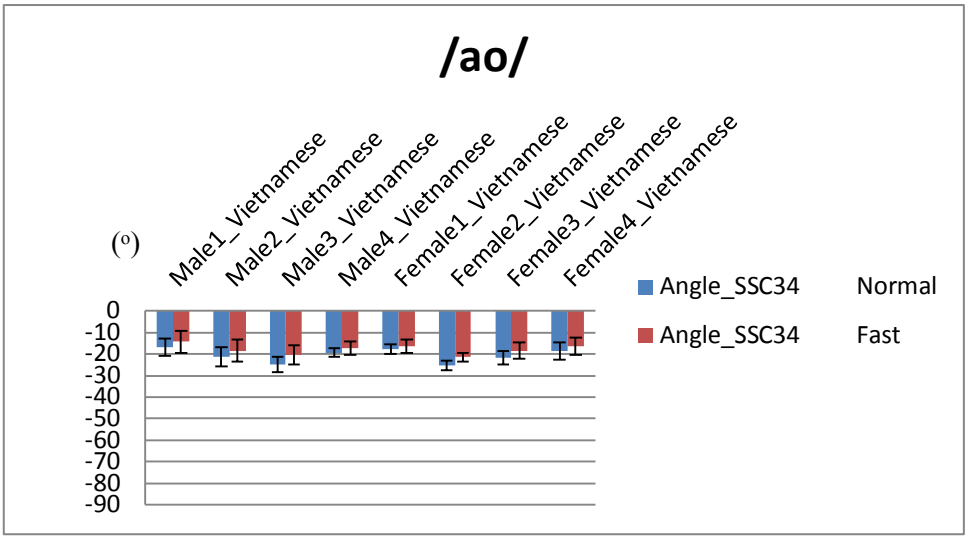


Figure A2-0-30: The average value and standard deviation of angle23 of /ao/ with all 8 Vietnamese subjects (4 males + 4 females).

A2-11, /aɔ/ sequence

❖ ***SSCF Angle12 of /aɔ/***

Figure A2-0-31 presents the average results of angle12 of the transition /aɔ/ of eight Vietnamese speakers (four males and four females).

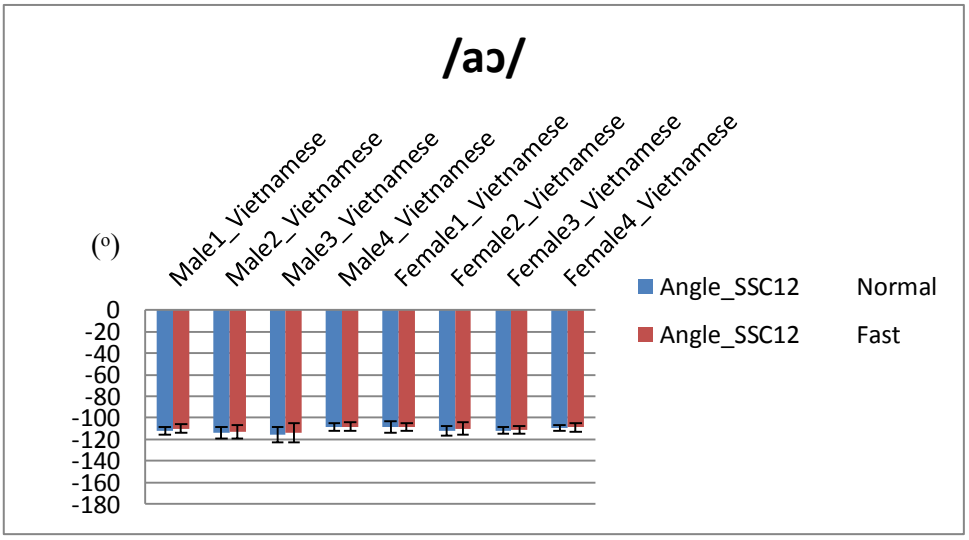


Figure A2-0-31: The average value and standard deviation of angle12 - /aɔ/ with all 8 Vietnamese subjects (4 males + 4 females).

Some comments are given, as following:

- For each subject, angle12 of transition /aɔ/ is more or less the same with both normal and fast rate with its small standard deviation.
- The angle12 of transition /aɔ/ among eight speakers are more or less the same value.

❖ ***SSCF Angle23 of /aɔ/***

Figure A2-0-32 presents the average results of angle23 of the transition /aɔ/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle23 of transition /aɔ/ is more or less the same with both normal and fast rate.
- The angle23 of transition /aɔ/ among eight speakers are more or less the same value.

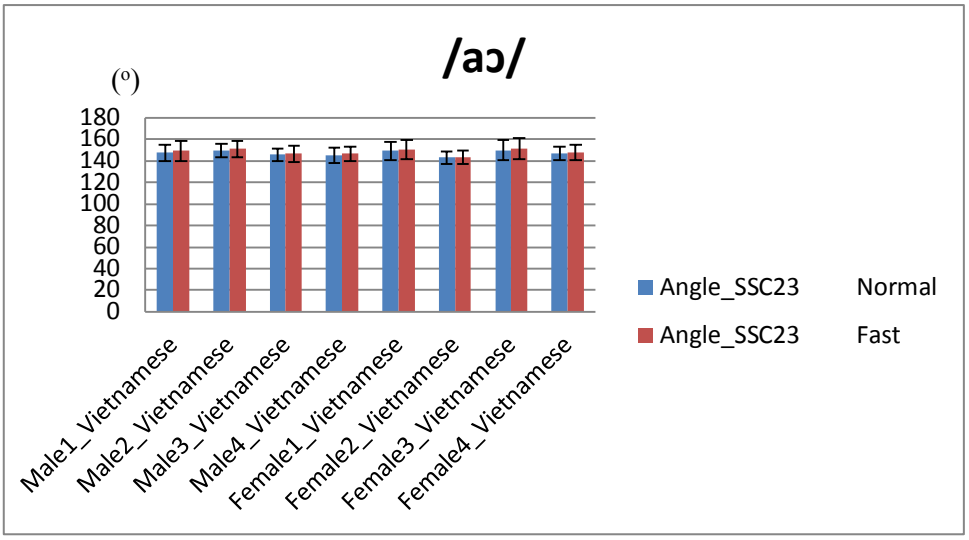


Figure A2-0-32: The average value and standard deviation of angle23 of /aɔ/ with all 8 Vietnamese subjects (4 males + 4 females).

❖ ***SSCF Angle34 of /aɔ/***

Figure A2-0-33 presents the average results of angle34 of the transition /aɔ/ of eight Vietnamese speakers (four males and four females). Some comments are given, as following:

- For each subject, angle34 of transition /aɔ/ is more or less the same with both normal and fast rate
- Among eight speakers, there are three groups:
  - ✓ First group: the angle34s of transition /aɔ/ of male1and feamle1 are more or less the same value.
  - ✓ Second group: the angle34s of transition /aɔ/ of Male1, Male3, Male4 and Female4 are more or less the same value.
  - ✓ Final group: the angle34s of transition /aɔ/ of Female2 and Female3 are more or less the same value.

- ✓ The angle34s of three groups are not the same value, but these differences here are small.

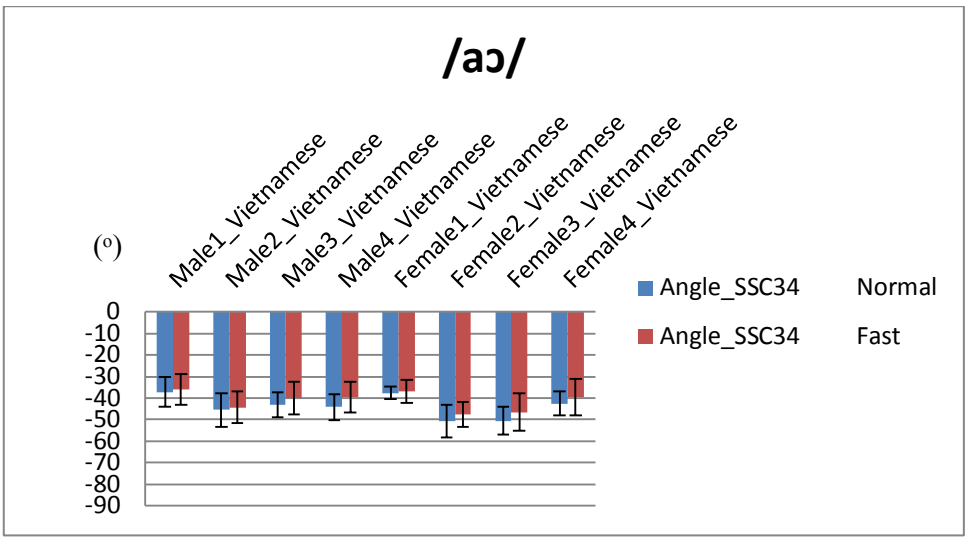


Figure A2-0-33: The average value and standard deviation of angle23 of /aɔ/ with all 8 Vietnamese subjects (4 males + 4 females).

Appendix 3. SSCF Angles comparisons among different French V1V2 transitions for each speaker

This part will show the comparison results of three angles (angle12, angle23, angle34) among the French different transitions (/ai, aɛ, ae, ua, ui/) for each speaker with two speech rates (normal and fast rate) (from 1 French male M2_Fr and 1 French female F2_Fr), as follows:

- **SSCF Angle12**

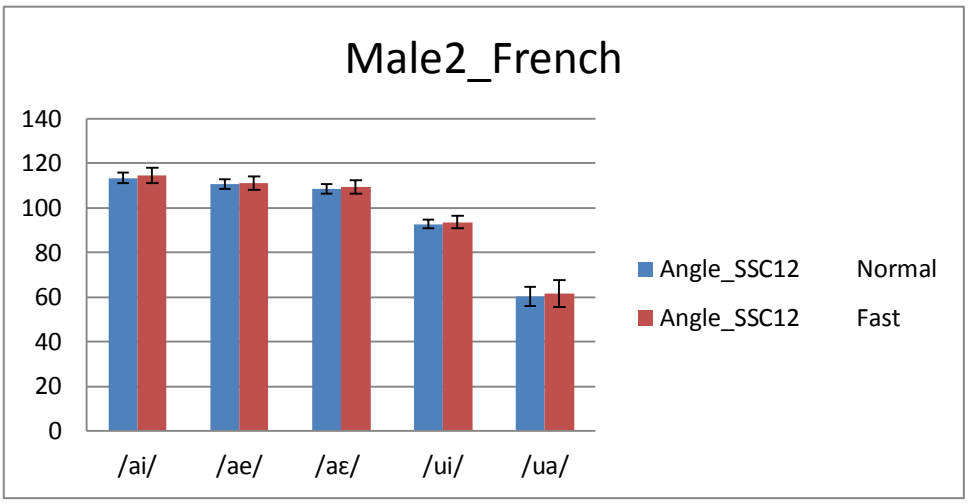


Figure A3-0-1: The average value and standard deviation of SSCF Angle12 of /ai, aɛ, ae, ua, ui/ transitions of French male2.

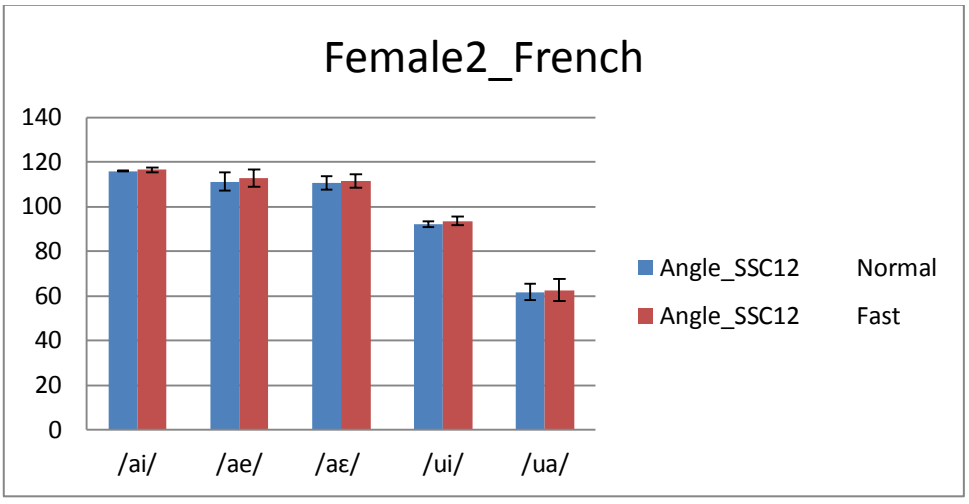


Figure A3-0-2: The average value and standard deviation of SSCF Angle12 of /ai, aɛ, ae, ua, ui/ transitions of French female2.

❖ SSCF Angle23

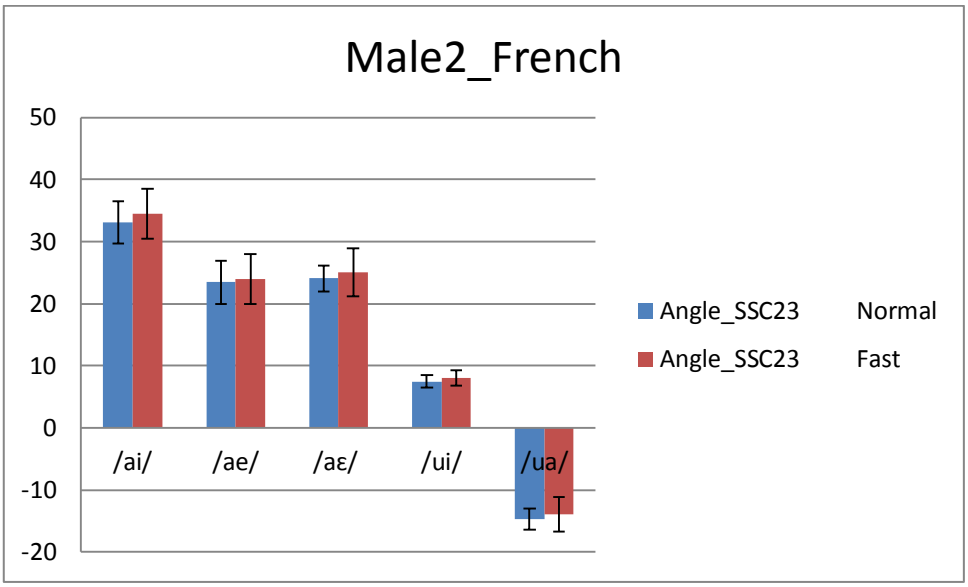


Figure A3-0-3: The average value and standard deviation of SSCF Angle23 of /ai, aɛ, ae, ua, ui/ transitions of French male2.

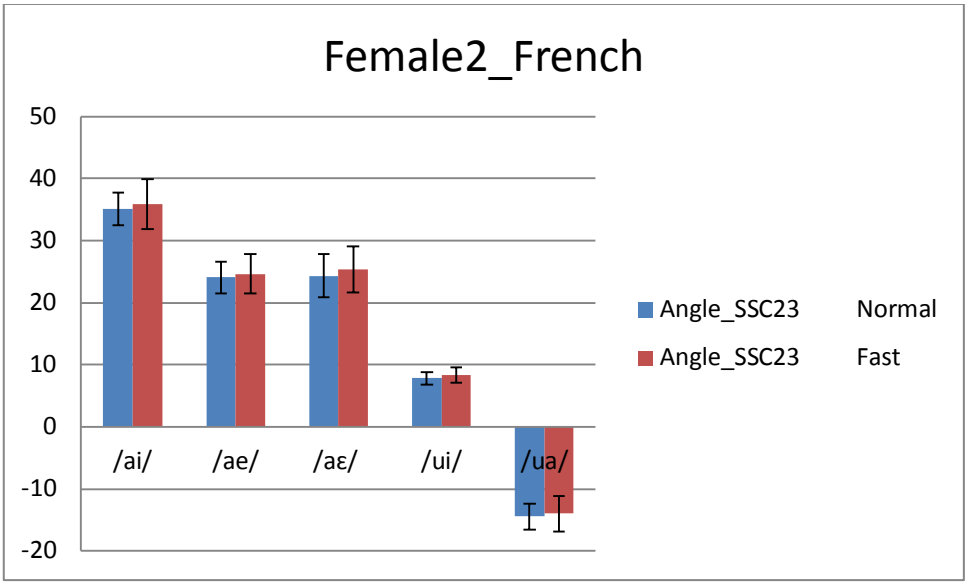


Figure A3-0-4: The average value and standard deviation of SSCF Angle23 of /ai, aɛ, ae, ua, ui/ transitions of French female2.

❖ SSCF Angle34

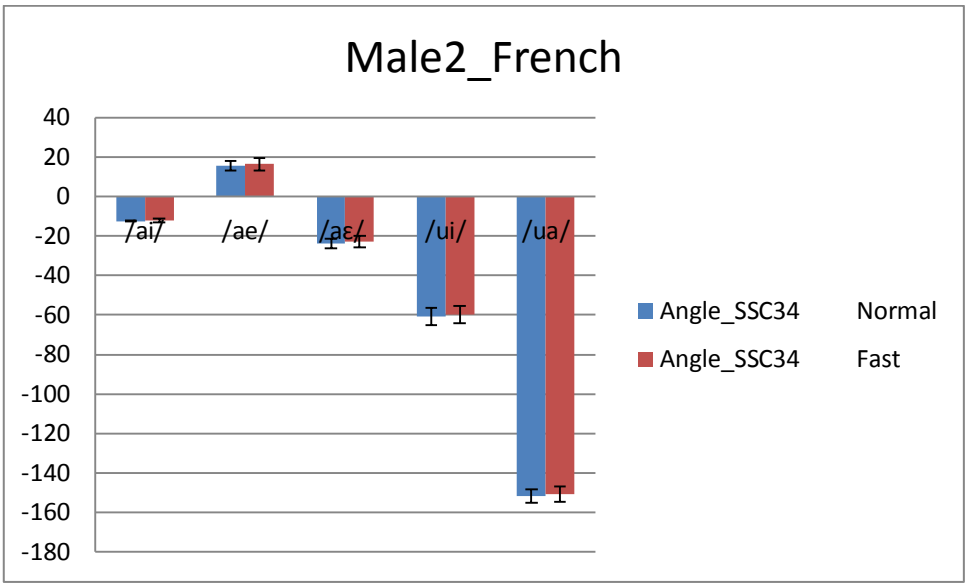


Figure A3-0-5: The average value and standard deviation of SSCF Angle34 of /ai, aɛ, ae, ua, ui/ transitions of French male2.

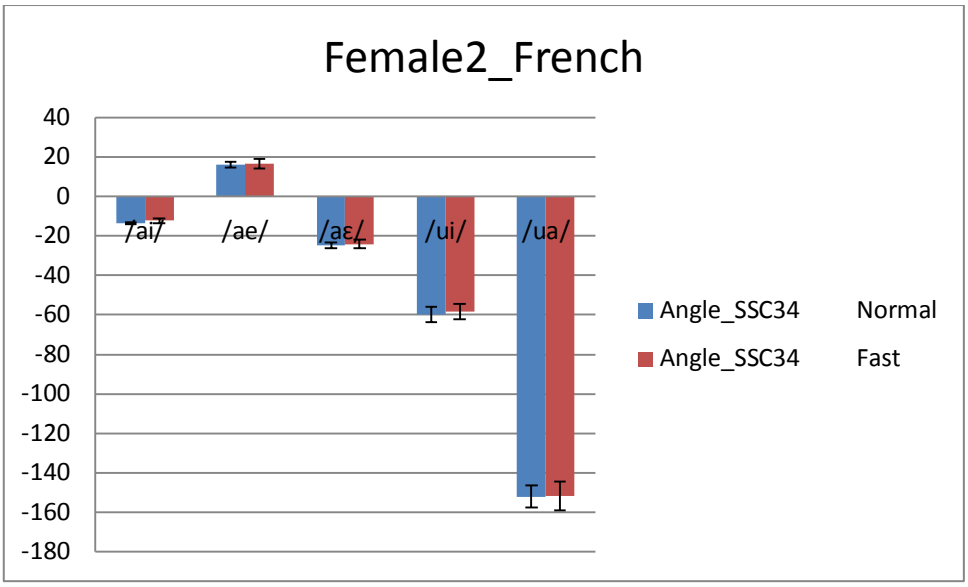


Figure A3-0-6: The average value and standard deviation of SSCF Angle34 of /ai, aɛ, ae, ua, ui/ transitions of French female2.

Appendix 4. SSCF Angles comparisons with same French V1V2 transitions produced by different speakers

This part will present the comparison of three SSCF Angles (angle12, angle23, angle34) of the same transition produced by four French native speakers (2 males + 2 females) at both normal and fast rate.

A4-1, /aɛ/ sequence

❖ **SSCF Angle12 of /aɛ/**

Figure A4-0-1 presents the average results of angle12 of the transition /aɛ/ of four French speakers (two males and two females). Some comments are given, as following:

- For each subject, angle12 of transition /aɛ/ is more or less the same with both normal and fast rate with its small standard deviation.
- The angle12 of transition /aɛ/ among four speakers are more or less the same value.

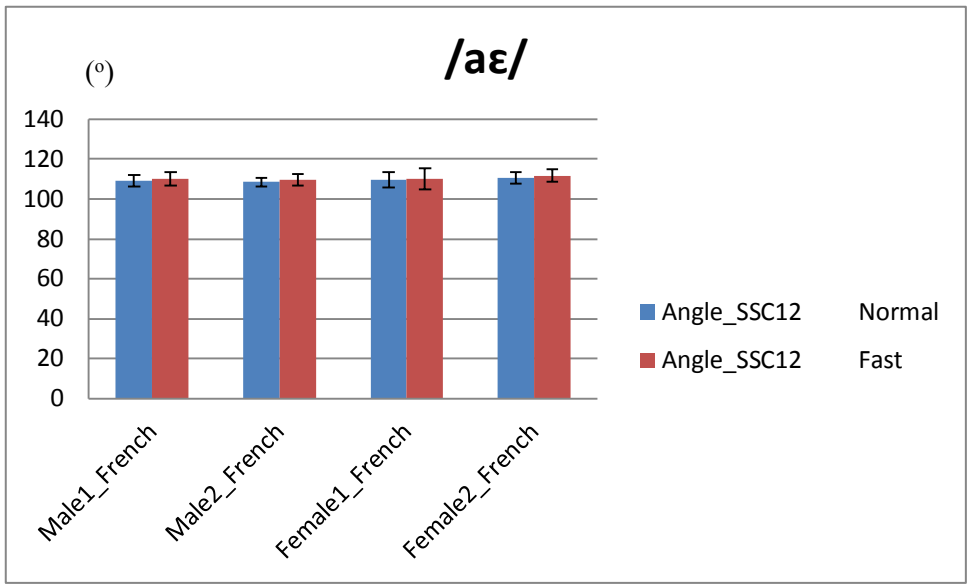


Figure A4-0-1: The average value and standard deviation of SSCF Angle12 of /aɛ/ with all 4 French Vietnamese participants (2 males + 2 females).

❖ **SSCF Angle23 of /aɛ/**

Figure A4-0-2 presents the average results of angle23 of the transition /aɛ/ of four French speakers (two males and two females). Some comments are given as following:

- For each subject, angle23 of transition /aɛ/ is more or less the same with both normal and fast rate with its small standard deviation.
- The angle23 of transition /aɛ/ among four speakers are more or less the same value.

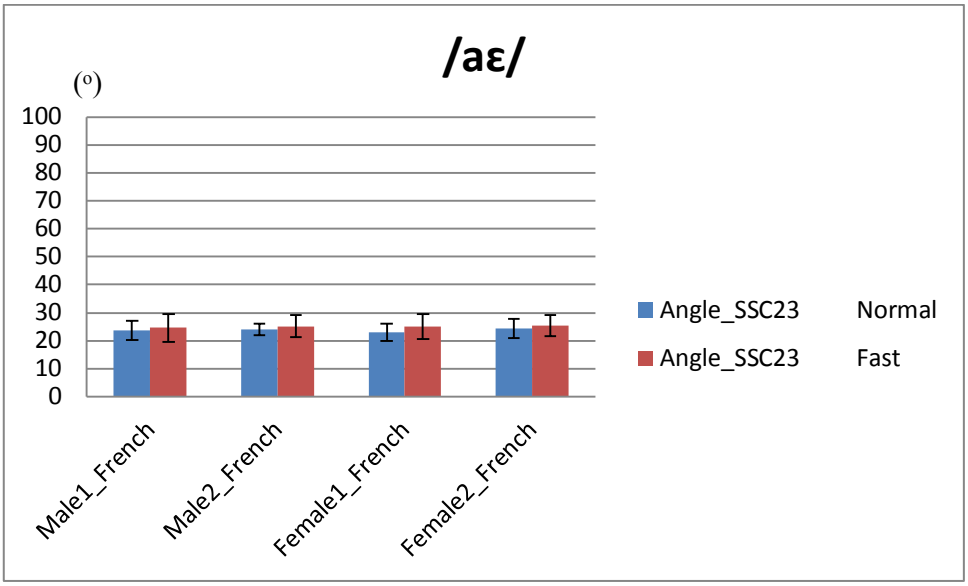


Figure A4-0-2: The average value and standard deviation of SSCF Angle12 of /aɛ/ with all 4 French Vietnamese participants (2 males + 2 females).

❖ **SSCF Angle34 of /aɛ/**

Figure A4-0-3 presents the average results of angle34 of the transition /aɛ/ of four French speakers (two males and two females). Some comments are given, as following:

- For each subject, angle34 of transition /aɛ/ is more or less the same with both normal and fast rate with its very small standard deviation.
- The angle34 of transition /aɛ/ among four speakers are more or less the same value.

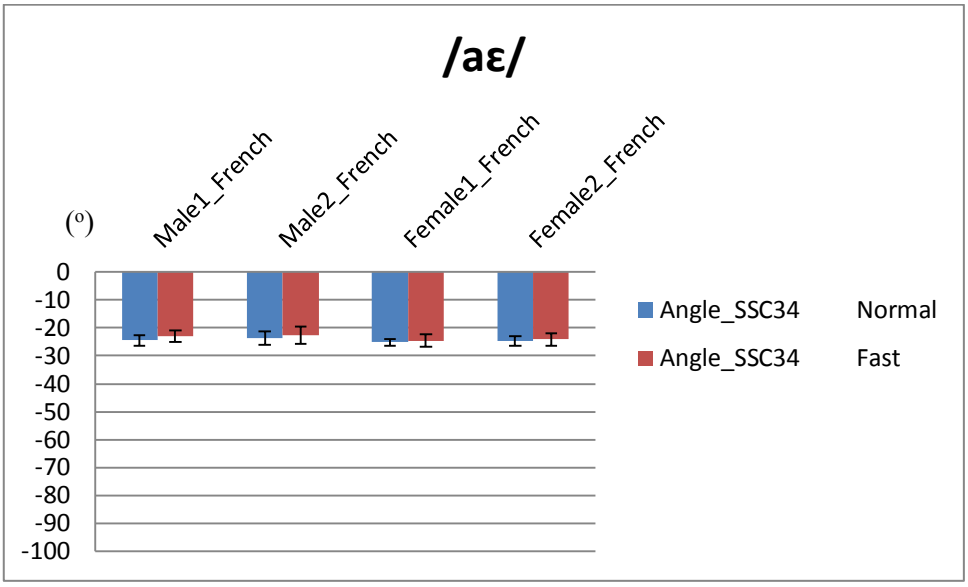


Figure A4-0-3: The average value and standard deviation of SSCF Angle 34 of /aɛ/ with all 4 French Vietnamese participants (2 males + 2 females).

A4-2, /ae/ sequence

❖ ***SSCF Angle12 of /ae/***

Figure A4-0-4 presents the average results of angle12 of the transition /ae/ of four French speakers (two males and two females). Some comments are given, as following:

- For each subject, angle12 of transition /ae/ is more or less the same with both normal and fast rate with its small standard deviation.
- The angle12 of transition /ae/ among four speakers are more or less the same value.

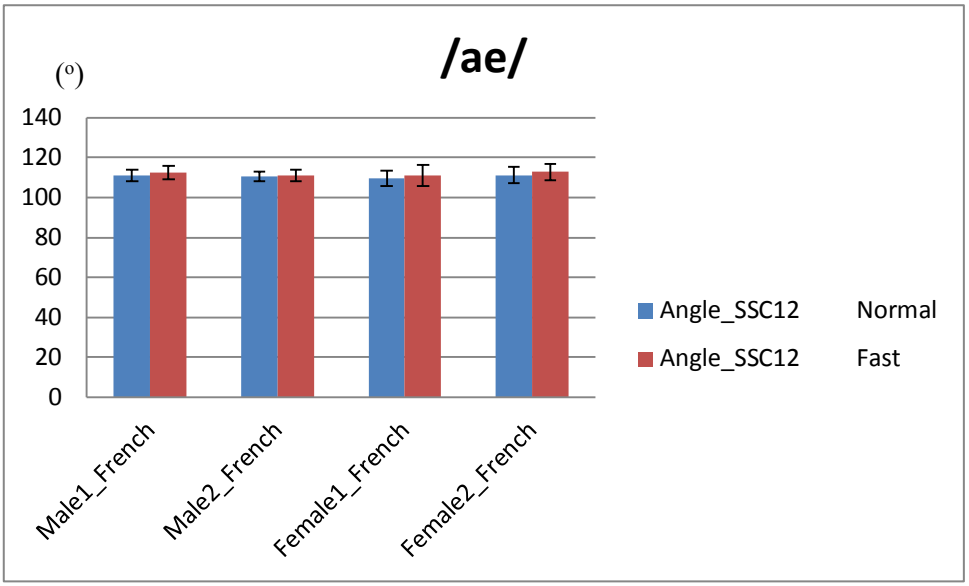


Figure A4-0-4: The average value and standard deviation of SSCF Angle12 of /ae/ with all 4 French Vietnamese participants (2 males + 2 females).

❖ ***SSCF Angle23 of /ae/***

Figure A4-0-5 presents the mean results of angle23 of the transition /ae/ of four French speakers (two males and two females). Some comments are given, as following:

- For each subject, angle23 of transition /ae/ is more or less the same with both normal and fast rate with its small standard deviation.
- Although the standard deviations of the angle23's transition /ae/ of four speakers are not very small, but as a whole, the angle23s of four speakers are more or less the same value.

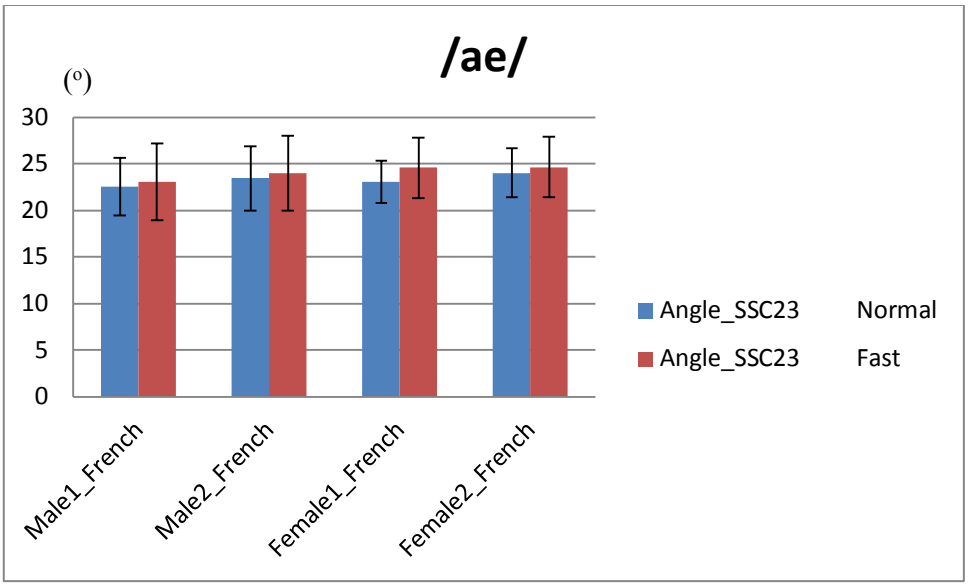


Figure A4-0-5: The average value and standard deviation of SSCF Angle23 of /ae/ with all 4 French Vietnamese participants (2 males + 2 females).

❖ ***SSCF Angle34 of /ae/***

Figure A4-0-6 presents the average results of angle34 of the transition /ae/ of four French speakers (two males and two females). Some comments are given, as following:

- For each subject, angle34 of transition /ae/ is more or less the same with both normal and fast rate.
- Although the standard deviations of the angle23's transition /ae/ of four speakers are not very small, but as a whole, the means of angle34 of four speakers are more or less the same value.

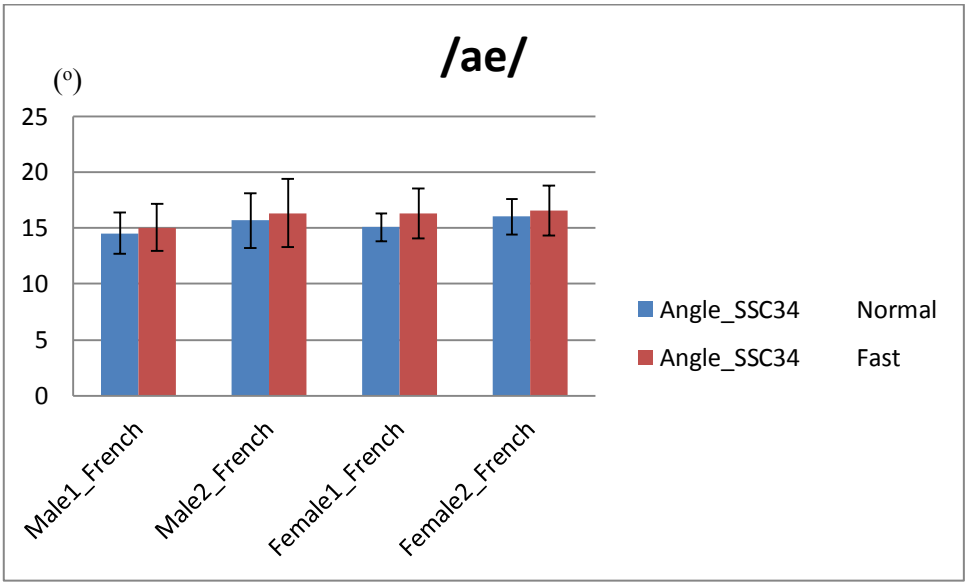


Figure A4-0-6: The average value and standard deviation of SSCF Angle23 of /ae/ with all 4 French Vietnamese participants (2 males + 2 females).

RESUME DE LA THESE EN FRANÇAIS

Résumé

La sélection de caractéristiques acoustiques appropriées est essentielle dans tout système de traitement de la parole. Pendant près de 40 ans, la parole a été généralement considérée comme une séquence de signaux quasi-stables (voyelles) séparés par des transitions (consonnes). Bien qu'un grand nombre d'études documentent clairement l'importance de la coarticulation, et révèlent que les cibles articulatoires et acoustiques ne sont pas indépendantes du contexte, l'hypothèse que chaque voyelle présente une cible acoustique qui peut être spécifiée d'une manière indépendante du contexte reste très répandue. Ce point de vue implique des limitations fortes. Il est bien connu que les fréquences de formants sont des caractéristiques acoustiques qui présentent un lien évident avec la production de la parole, et qui peuvent participer à la distinction perceptive entre les voyelles. Par conséquent, les voyelles sont généralement décrites avec des configurations articulatoires statiques représentées par des cibles dans l'espace acoustique, généralement par les fréquences des formants correspondants, représentées dans les plans F1-F2 et F2-F3. Les consonnes occlusives peuvent être décrites en termes de point d'articulation, représentés par locus (ou locus équations) dans le plan acoustique. Mais les trajectoires des fréquences de formants dans la parole fluide présentent rarement un état d'équilibre pour chaque voyelle. Elles varient avec le locuteur, l'environnement consonantique (co-articulation) et le débit de parole (relative à un continuum entre hypo et hyper-articulation). En vue des limites inhérentes aux approches statiques, la démarche adoptée ici consiste à étudier les transitions entre les voyelles et les consonnes (V1V2 et V1CV2) d'un point de vue dynamique.

Tout d'abord, nous avons étudié les effets de la réponse impulsionnelle en début, à la fin et pendant les transitions du signal, à la fois dans le signal de parole et au niveau de la perception. Les variations des phases des composantes ont ensuite été examinées. Les résultats montrent que les effets de ces paramètres peuvent être observés dans les spectrogrammes. Fondamentalement, les amplitudes des composantes spectrales qui se distinguent dans l'approche préconisée ici sont suffisantes pour une discrimination perceptive. De ce fait, pour toute analyse de la parole, nous nous concentrons uniquement sur le domaine de l'amplitude, laissant délibérément de côté les informations de phase. Ensuite, nous avons mesuré la perception des transitions voyelle-consonne-voyelle d'un point de vue dynamique. Ces résultats perceptifs, avec ceux obtenus précédemment par Carré (2009a), montrent que les stimuli des transitions voyelle-voyelle et voyelle-consonne-voyelle peuvent être caractérisés et séparés par la direction et la vitesse des transitions formantiques, même lorsque les valeurs absolues de la fréquence sont en dehors du triangle vocalique (c'est-à-dire en dehors de l'espace acoustique de la voyelle en valeur absolue). En raison des limites intrinsèques à la mesure des formants, l'approche dynamique a besoin de développer de nouveaux outils, pour définir des paramètres qui peuvent

remplacer l'estimation de la fréquence des formants. C'est pourquoi les values fréquentielle des centroïde de sous-bande spectrales (Spectral Sub-band Centroïde Frequency) ont été mesurées et étudiées (SSCF). Des comparaisons avec les fréquences de formants des voyelles montrent que les paramètres SSCFs peuvent remplacer l'estimation des fréquences des formants et ainsi d'agir comme des « pseudo-formants », même pendant la production des consonnes.

Sur cette base, les paramètres SSCF sont utilisés comme un outil pour calculer les caractéristiques dynamiques. Nous proposons une nouvelle façon de modéliser les caractéristiques dynamiques de la parole: nous l'avons appelé *SSCF Angles*. Notre travail d'analyse sur les *SSCF Angles* a été réalisé sur les transitions de séquences vocaliques V1V2 prononcées par des Vietnamiens et des Français. Les *SSCF Angles* apparaissent comme des paramètres fiables et robustes. Pour chaque langue, les résultats de l'analyse montrent que: (1) les *SSCF Angles* peuvent permettre la distinction entre transitions vocaliques de type V1V2 ; (2) les transitions V1V2 et V2V1 montrent des propriétés symétriques dans le domaine acoustique sur cette base de paramètres *SSCF Angles* ; (3) les *SSCF Angles* pour les hommes et les femmes sont assez similaires pour une même transition étudiée de contexte V1V2 ; et (4) ils sont aussi plus ou moins invariants en fonction du débit de parole (débit de parole normal et rapide). Enfin, ces caractéristiques vocales acoustiques dynamiques sont utilisées dans un système automatique de reconnaissance vocale en langue vietnamienne.

Mots clés: gestes vocaliques, caractéristiques dynamiques acoustiques, amplitude, direction et vitesse de la transition, SSCF Angles, reconnaissance automatique de la parole.

Introduction

La parole joue un rôle essentiel dans la communication humaine. Aujourd'hui, avec les progrès de l'information et de l'électronique, de nombreuses machines sophistiquées sont utilisées dans la vie quotidienne, et les individus veulent être en mesure de communiquer avec ces machines par la voix. La communication orale est utile pour les personnes quand leurs mains sont occupées; elle est également cruciale pour les personnes aveugles. Nous pouvons trouver de nombreux domaines d'applications dans la vie réelle, tels que des interfaces utilisateur utilisant la voix, la numérotation vocale (par exemple, "home Call"), le routage des appels (par exemple: « Je voudrais faire un appel en PCV »), le contrôle de l'électroménager, la recherche d'informations (par exemple, trouver un podcast où des mots particuliers ont été prononcés), la saisie simple de données (par exemple, la saisie d'un numéro de carte de crédit), la préparation de documents structurés (par exemple, un rapport de radiologie fait par un médecin), et le traitement de la parole en texte (par exemple, le traitement de texte ou e-mails), etc.

Le traitement de la parole est l'étude des signaux de parole et les méthodes de traitement des signaux. Les signaux sont maintenant généralement traités dans une représentation numérique, de sorte que le traitement de la parole peut être considéré comme un cas particulier du traitement numérique du

signal, appliqué au signal de parole. La synthèse vocale et la reconnaissance vocale sont deux grandes parties du traitement de la parole. La synthèse vocale est la production artificielle de la parole humaine. Ce système convertit le texte en langage oral. Et à l'inverse, un système de reconnaissance automatique de la parole (RAP) convertit les mots parlés en texte : l'entrée d'un système RAP est le signal de parole et la sortie est la séquence de texte correspondant à ce signal.

La sélection des caractéristiques acoustiques appropriées est peut-être la tâche la plus importante dans la conception d'un système qui utilise le traitement de la parole.

Les humains sont capables de comprendre les voix d'enfants ainsi que les voix des adultes, les voix masculines ainsi que les voix féminines, mais les systèmes RAP actuels ne disposent pas de cette capacité. Cette lacune provient des limites inhérentes à la méthode de comptabilisation fondée sur la modélisation statistique des propriétés spectrales.

Pendant environ 40 ans, une globalisation des théories de traitement automatique de la parole a considéré la parole comme des séquences de voyelles et de consonnes en tant que parties du discours. Dans certaines langues du monde, un mot peut être formé sans consonnes, par contre aucun mot ne peut exister qu'avec des consonnes sans une voyelle. L'hypothèse largement utilisée est que chaque voyelle constitue une partie stable appelée « cible acoustique ». Dans l'approche classique du système de production de la parole humaine, ces cibles acoustiques doivent être atteintes lors de la production des voyelles. Par conséquent, les voyelles jouent un rôle très important dans la façon dont chaque mot est reconnu. Mais l'arrangement de la parole en phonèmes successifs à un niveau phonologique ne signifie nullement que, d'un point de vue phonétique, la parole est faite d'une succession d'états stables. (Greenberg et al., 2002).

Plus récemment, dans la seconde moitié du XX^{e} siècle, certains chercheurs ont démontré que la parole naturelle est de nature dynamique et non pas une simple séquence de segments à l'état stable. Certains résultats sont alors proposés pour la production dynamique de la parole et de la perception par son application (voir par example Strange et al, 1983;. W Strange, 1989a; Divenyi et al., 2006; Carré, 2009a; Hermansky, 2011). Des résultats intéressants obtenus par Carré (2009a) sur la transition voyelle à voyelle, montrent de manière surprenante que, même lorsque les deux voyelles cibles ne sont pas réellement atteintes, le son produit est entendu de la même manière que lorsque les cibles sont réalisées, tant que la direction et la pente de la transition sont maintenues. Cette caractéristique a été appelé « geste vocalique » ou « geste acoustique » étendu.

À la suite de ce résultat, il devient possible de modéliser l'évolution dynamique du signal de parole au fil du temps comme une série de « gestes acoustiques », et cela pourrait être appliqué dans le système de production de la parole ainsi que dans les systèmes RAP. Dans le système de production de parole, la cible peut ne pas avoir besoin d'être atteinte. Les « gestes acoustiques » peuvent être suffisants pour indiquer la cible visée. Dans un système RAP, nous aurons la définition d'un critère

discriminant qui est pour chaque geste identifié la définition d'une mesure capable de le reconnaître comme distinct de tous les autres.

L'objectif de la thèse

Par conséquent, nous proposons cette thèse pour essayer de développer un type de caractéristiques vocales acoustiques dynamiques dans le domaine acoustique (et non pas dans les domaines articulatoire et/ou perceptif) afin qu'ils puissent satisfaire aux impératifs suivants :

- Des mesures « fiables » sur le signal ;
- des paramètres indépendants du locuteur (et donc du genre);
- utilisable pour reconnaissance automatique de la parole RAP.

Contexte de l'étude

Dans notre thèse, nous nous concentrons sur l'analyse des caractéristiques des deux langues: l'une est une langue tonale – le vietnamien (notre langue maternelle) et l'autre est une langue étrangère non-tonale – le français.

Le vietnamien est une langue austroasiatique, c'est la langue nationale officielle du Vietnam. C'est la langue maternelle du peuple vietnamien (Kinh), ainsi que la première ou deuxième langue pour de nombreuses minorités ethniques du Vietnam (Wiki-anglais, 2015). Le vietnamien est une langue isolante et une langue tonale (Trương, 1970; Nathan et Cao, 1988; Mai et al., 1997; Đoàn, 1999). Les mots vietnamiens sont constitués d'une ou plusieurs syllabes (Truong, 1970; Đoàn, 1999). Le vietnamien dispose d'un système tonal complexe défini non seulement par la modulation de hauteur dans la syllabe, mais aussi par les caractéristiques de qualité de la voix (tonals glottalisés) (Michaud, 2005, 2010; Brunelle, 2009). Le nombre de tons vietnamiens peut varier de six (variété du nord) à cinq (variété du sud), ou quatre dans certains dialectes du centre (Nguyen, 2004). Chaque syllabe vietnamienne porte un unique ton. A la différence du vietnamien, le français est une langue non-tonale.

Ces deux langues ont des différences dans les structures possibles de syllabes (Trần, 2011). Il y a aussi des similitudes et des différences dans le système de consonnes et de voyelles entre les deux langues (Trần, 2011). Dans le cadre de notre travail de thèse, nous allons aborder les comparaisons entre les deux langues pour des voyelles /a, ɛ, e, i, u/ similaires aux deux langues sur la base de la caractéristique dynamique acoustique proposée de la parole qui sera présenté dans ce thèse.

Tâche principale

La première tâche que nous avons menée et présentée Chapitre 3 a consisté tout d'abord en l'extraction de paramètres acoustiques pertinents, qui pourraient être des résonances de fréquence ou des pics de spectre de fréquence ou leurs pseudo-paramètres, sans prendre en compte les caractéristiques de la perception humaine.

Puis nous essayons dans la suite du Chapitre 3 et au Chapitre 4 d'extraire les caractéristiques des « formes » et leurs représentations pour les phonèmes et aussi pour les transitions (voyelle-voyelle, voyelle-consonne ou consonne-voyelle).

Enfin, nous avons intégré ces fonctionnalités dans un système automatique de reconnaissance vocale en langue vietnamienne pour évaluer l'effet de ces modèles de caractéristiques dans le Chapitre 5.

Le contenu de la thèse

(Toutes les figures et les tableaux correspondent aux ordres de figures et de tableaux dans le manuscrit original en anglais)

Les caractéristiques dynamiques de la parole acoustique sont le fil conducteur de cette thèse, dans laquelle elles sont étudiées dans le cadre de la perception de la parole, l'analyse de la parole, et la reconnaissance de la parole pour les Vietnamiens.

Chapitre 1 : l'état de l'art sur le traitement de la parole fait ressortir les nombreuses limites d'une vision statique de la parole. Des travaux de recherche menés dans la seconde moitié du XX^{e} siècle ont démontré que la parole est plutôt un processus dynamique dans la production et perception de la parole, ces travaux ont ouvert une série de nouvelles voies et axes de recherche, auxquels cette thèse apporte une contribution supplémentaire.

Chapitre 2 : Étude de certaines caractéristiques de la parole sur les deux aspects statiques et dynamiques, à la fois dans le signal de parole et au niveau de la perception. Deux expériences ont été réalisées afin de tester les effets de l'impulsion au début, à la fin, et au cours de la transition des signaux de parole. Les résultats obtenus ont montré que ces effets sont faibles – ce qui vient comme de bonnes nouvelles dans la perspective d'une recherche de corrélats acoustiques stables de segments de parole. Par ailleurs, le signal de parole est entièrement caractérisé par son spectre de puissance et le spectre de phase. Par conséquent, nous en sommes venus à étudier le rôle du spectre de phase dans le test de perception, spectre qui est considéré alors comme un bon candidat pour caractériser la dynamique de la parole. Les résultats ont montré que la perception des amplitudes des composantes spectrales est nettement suffisante pour que la discrimination soit révélée sans information du spectre de phase.

Compte tenu des résultats obtenus dans le Chapitre 2, les travaux des chapitres suivants consistent à trouver un autre candidat dans le domaine de l'amplitude comme caractéristique dynamique du signal de parole (sans spectre de phase) qui peut être utile pour la discrimination de la parole en perception.

Dans le **Chapitre 3**, nous avons étudié le rôle des transitions formants pour la perception de séquences de parole V1CV2. Les résultats obtenus ont montré qu'une bonne perception des séquences V1CV2 synthétisées peut être obtenue avec une transition de fréquence de formants située à l'intérieur

et à l'extérieur du triangle vocalique. Ces résultats ont permis d'étendre de précédents résultats publiés dans l'approche déductive de Carré (2004), qui propose une vision dynamique de la production de la parole, et sur la prédiction des systèmes vocaliques de Carré (2007, 2009a, 2009b) et Nguyen (2009): la direction de la transition et la durée (pente) dans le domaine temporel peuvent permettre de distinguer les transitions VV, et les stimuli CV. Cependant, tous ces paramètres sont basés sur une analyse de variations des fréquences des formants, de sorte qu'ils souffrent des limitations inhérentes à l'estimation des fréquences des formants. C'est pourquoi notre travail décrit au **Chapitre 4** vise à trouver une autre caractéristique de la parole que les fréquences des formants. Les paramètres choisis sont les *Spectral Subband Centroid Frequency* (SSCF). Certains travaux d'analyse ont été effectués pour montrer que les paramètres SSCF sont semblables à des fréquences de formants mais avec l'avantage que ces paramètres restent continus, même pendant la production de consonnes contrairement aux fréquences formants.

Les paramètres *Spectral Subband Centroid* (SSC) ont été proposés par Paliwal (1998) et définis comme suit:

- Etape 1: diviser la bande des fréquences (0 à Fs / 2 où Fs est la fréquence d'échantillonnage en Hz) en un nombre déterminé de M sous-bandes. Chaque sous-bande a des bords inférieurs et supérieurs et une forme de filtre;
- Etape 2: Calculer le centre de gravité pour chaque sous-bande en utilisant le spectre du signal vocal. Chaque centroïde est caractérisé alors par sa fréquence et son amplitude.

Chacun des paramètres $SSCF_m$ est calculé par la formule suivante (4-1):

$$SSCF_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^{\gamma}(f) df}{\int_{l_m}^{h_m} w_m(f) P^{\gamma}(f) df} \quad \textbf{(4-1)}$$

où l_m et h_m sont les bords inférieurs et supérieurs de la sous-bande mth, et respectivement; $w_m(f)$ est sa forme; P(f) est le spectre de puissance, et γ est une constante de commande de la plage dynamique du spectre de puissance.

SSCF est considéré comme un paramètre similaire à un formant (Paliwal, 1998), et ses caractéristiques peuvent être extraites facilement et de manière fiable à partir du spectre de puissance du signal de parole.

Le schéma de l'algorithme qui calcule les paramètres SSCF est présenté à la figure 4-2 ci-dessous. Cet algorithme est très simple.

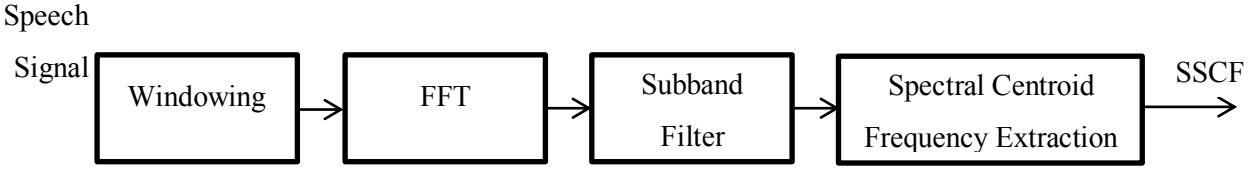


Figure 4-1: SSCF extraction algorithm.

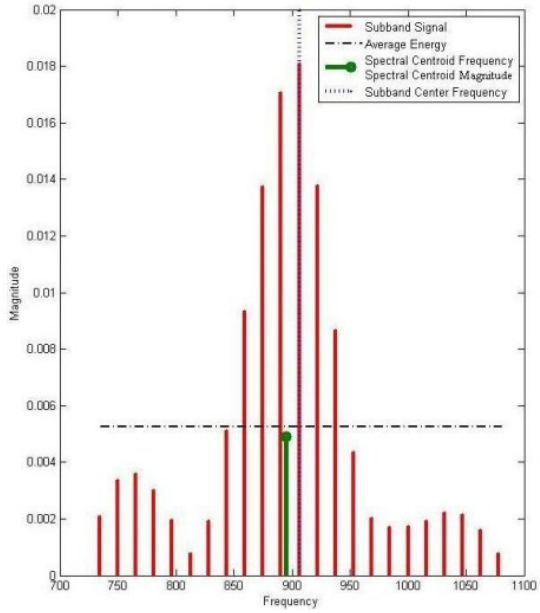


Figure 4-2: Subband signal, average energy (black dashed line), spectral subband centroid frequency (SSCF) and spectral subband centroid magnitude (SSCM).

Kuldip K. Paliwal - qui a proposé les caractéristiques de la SSC - a choisi de définir le nombre M = 3 de filtres de sous-bande parce que dans le cadre de ses travaux, ces paramètres n'ont été utilisés uniquement que comme un complément pour la reconnaissance vocale (Paliwal, 1998).

Selon notre point de vue, nous souhaitons développer une approche dynamique qui n'est pas seulement une approche statique *améliorée* avec des paramètres dynamiques, mais qui consiste en une approche intrinsèquement dynamique. Par conséquent, nous avons besoin de plus de trois paramètres SSC pour extraire plus d'informations sur le signal de parole pour modéliser les paramètres dynamiques. Ainsi, nous choisissons un nombre M = 6 de filtres de sous-bande à (correspondant à six pseudo-formants) et la banque de filtre est conçue en divisant la bande de fréquences de manière uniforme à l'échelle de mel avec une forme de triangle, comme dans la Figure 4-3.

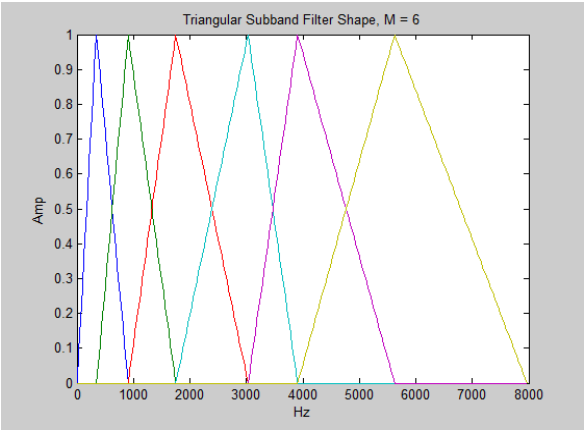
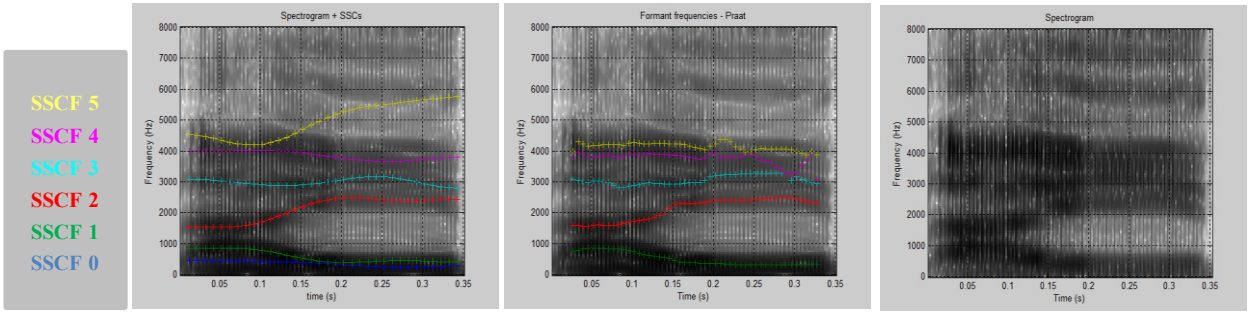


Figure 4-3: Subband filter shapes for computing SSCF with M = 6.

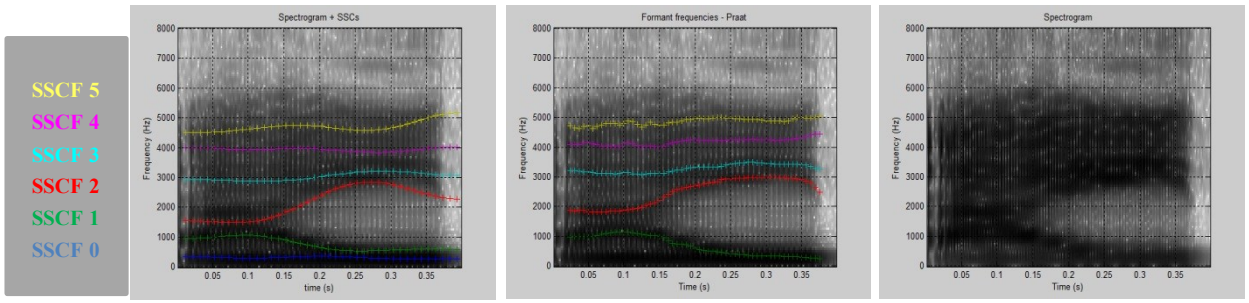
Caractéristiques SSCF aux propriétés similaires à des fréquences de formants

Les Figures 4-4 et 4-5 montrent un exemple de stimuli /ai/ prononcé en vietnamien respectivement par un locuteur natif et une locutrice native.



a) b) c)

Figure 4-4: /ai/ produced by one Vietnamese native male speaker: a) SSCF parameters (left); b) Formant frequencies (obtained from Praat toolkit); c) Spectrogram (right).



a) b) c)

Figure 4-5: /ai/ produced by one Vietnamese female speaker: a) SSCF parameters (left); b) Formant frequencies (obtained from Praat toolkit); c) Spectrogram (right).

En observant la Figure 4-4 pour la voix masculine, et en comparant les formes entre les paramètres SSCF (figure 4-4-a), les fréquences de formants (Figure 4-4-b) et les énergies dans le spectrogramme (Figure 4-4-c), nous pouvons voir clairement que pour ce stimuli/ ai /, les paramètres SSCF ont une forme similaire avec celle avec des fréquences de formants, en particulier les paramètres SSCS1 et SSF2 suivent de très près les évolutions des formants F1, F2,. Ce constat est aussi valable pour la voix féminine présentée la figure 4-5.

Paramètres SSCF continus sur le domaine temporel, à la différence des fréquences de formants

Pendant la production des consonnes, étant donné que les cordes vocales ne vibrent pas, et parce qu'un geste articulatoire de fermeture du conduit vocal est produit, il est difficile d'obtenir des valeurs et des évolutions fiables des fréquences des formants. Ces valeurs sont d'ailleurs généralement discontinues dans le domaine temporel du signal de la parole réelle, pendant la production des consonnes.

Au contraire, les paramètres SSCF ne sont évalués que sur le spectre de puissance du signal de parole ; il est donc facile de calculer les paramètres SSF même pendant la production de consonnes. Les évolutions de ces paramètres SSCF peuvent être alors représentées par des courbes continues dans

le domaine temporel. Nous voyons clairement ce point lors de l'observation des paramètres SSCF et des fréquences des formants sur les deux mêmes exemples de fricatives: les consonnes /s/ et /ʃ/ dans les stimuli /asi/ en Figure 4-8 et /aʃa/ dans la Figure 4-9 qui sont prononcées par un locuteur femme native vietnamienne.

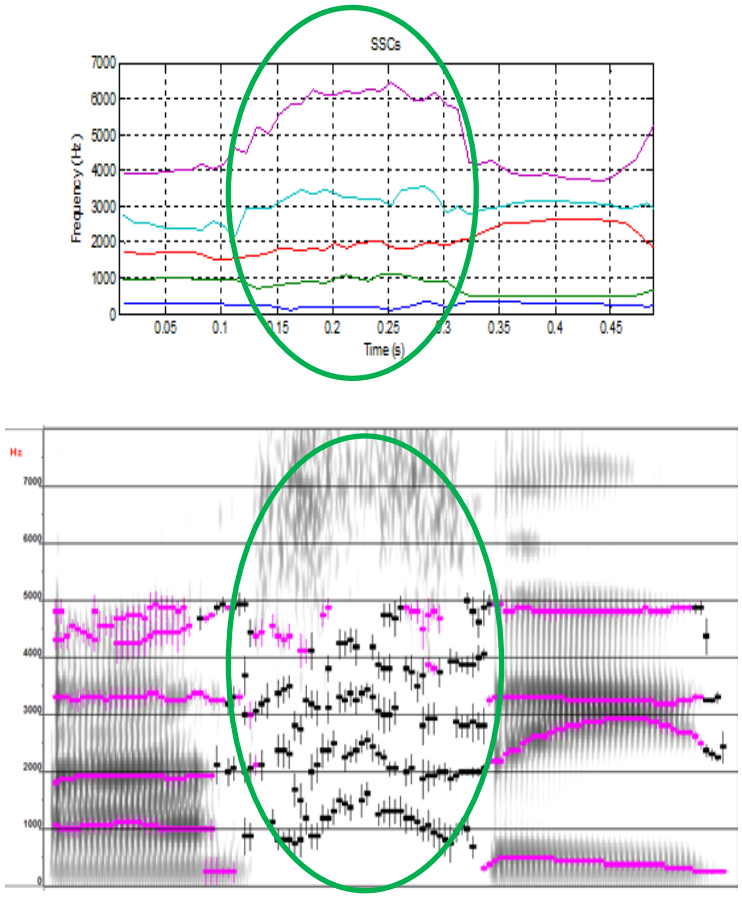


Figure 4-8: /asi/ of Vietnamese females: a*) SSCF parameters (top); b) formant frequencies (bottom: obtained from WinSnoori toolkit).*

La Figure 4-8 montre ainsi les évolutions pendant la production d'une fricatives /s/ non voisée alvéolaire prononcée par une femme vietnamienne native dans le cas du stimulus /asi/ : elles sont soulignées dans la région de l'ellipse verte. On constate que l'absence de nombreux points de mesure de fréquences de formants acoustique lors de la réalisation des fricatives /s/ est clairement visible et explique que les courbes de « suivi » des formants sont discontinues, tandis que les paramètres SSCF présentent quant à eux des courbes d'évolution continues pouvant être extraites au cours de cette production consonantique.

De même la Figure 4-9 souligne par l'ellipse verte la fricative alvéopalatale /ʃ/ produit par une locutrice femme vietnamienne native où sont marquées dans le cas du stimulus /aʃa/. Comme observé pour les fricatives /s/, les paramètres SSCF présentent des courbes d'évolution continues, même pendant la production de consonne /ʃ/, alors qu'il est difficile d'extraire les fréquences de formants.

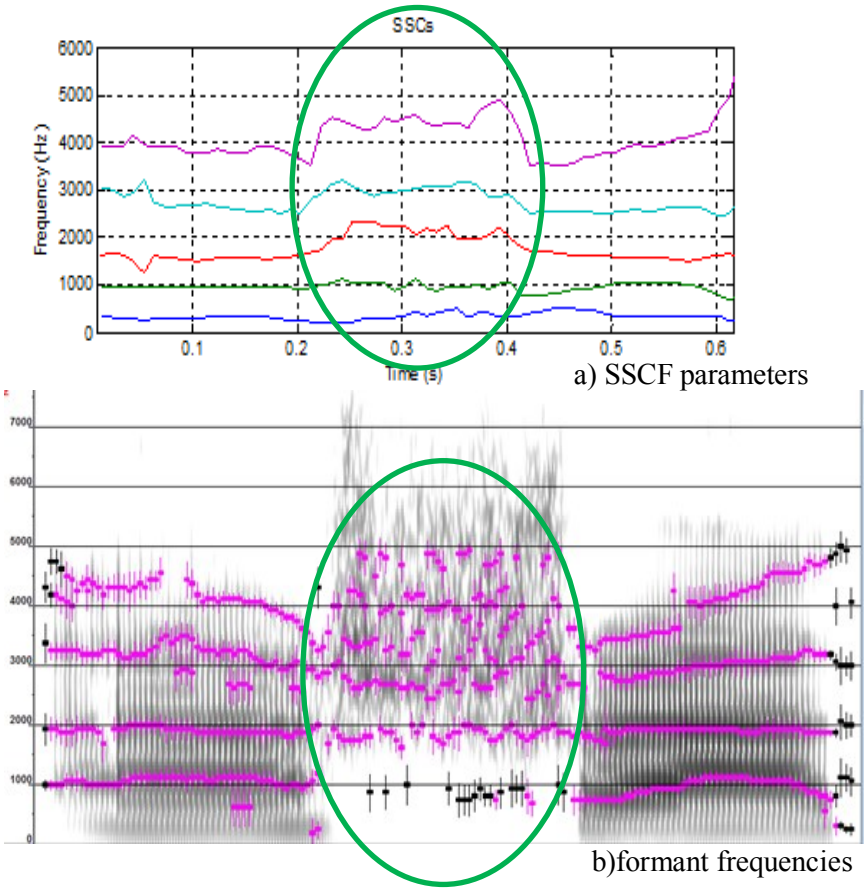


Figure 4-9: /aʃa/ of Vietnamese female: a) SSCF parameters (top); b) formant frequencies (bottom: obtained from WinSnoori toolkit).

De toute évidence, au contraire des fréquences de formants, les paramètres SSCF sont des paramètres continus dans le domaine temporel, même pendant la production de consonnes.

Modélisation acoustiques caractéristiques vocales dynamiques - SSCF Angles

Nous avons étudié six stimuli vietnamiens de type transition de voyelle à voyelle /ai, ea, au, ua, ui, iu/. Chaque élément a été enregistré dix fois (5 fois au taux normal et 5 fois à vitesse rapide) par des locuteurs vietnamiens natifs (1 homme et 1 femme) du nord du Vietnam avec un âge compris entre 25 et 32 ans.

Les figures 4-13 et 4-14 montrent les résultats graphiques de SSCF1/SSCF2 en fonction de leur vitesse respectivement pour un locuteur mâle vietnamien natif et pour une locutrice femme vietnamienne native.

L'analyse des résultats tracés dans ces deux Figures 4-13 et 4-14, permet de constater que d'une part, sur le plan SSCF1/SSCF2, toutes les trajectoires SSCF sont des lignes assez droites. Et trois paires de transition peuvent être évidemment séparées (sur la gauche de la Figure 4-13 pour la voix masculine vietnamienne, et à gauche de la Figure 4-14 pour la voix féminine vietnamienne): la première paire est un groupe de transitions /ai, ia/ qui sont modélisées respectivement par des lignes pleines ou en «+ » bleues ; la seconde paire est le groupe de transitions /au, ua/ qui sont modélisées

respectivement par des lignes solides ou en « + » vertes ; et une troisième paire est le groupe de /ui, iu/ qui sont modélisées respectivement par des lignes pleines ou « + » rouges.

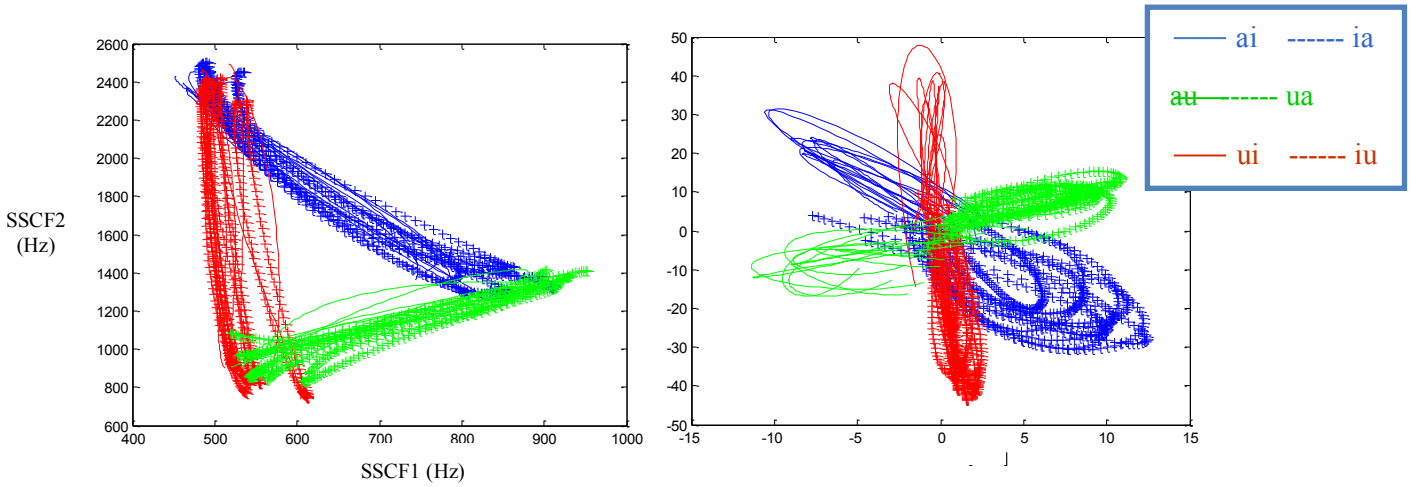


Figure 4-73: Vowel-to-Vowel transitions on SSCF1/SSCF2 plane and transition speeds produced by a native male speaker of Vietnamese.

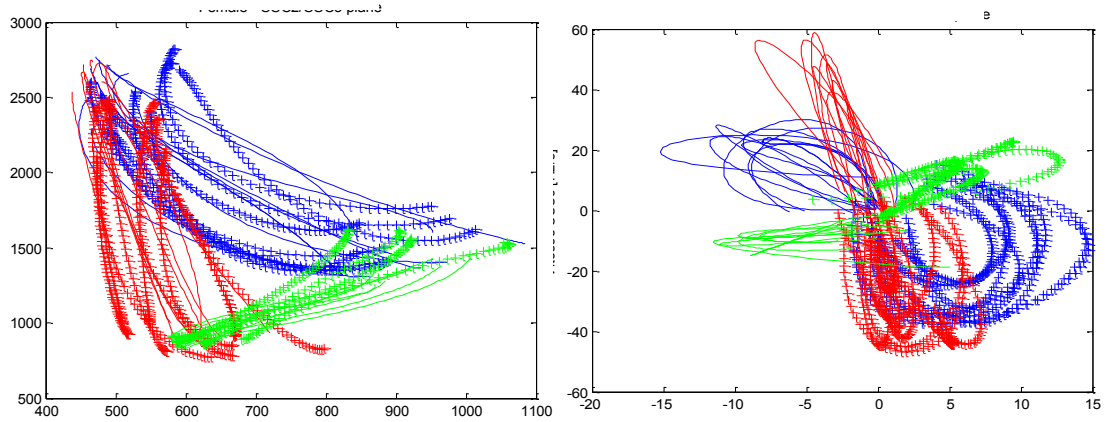


Figure 4-84: Vowel-to-Vowel transitions on SSCF1/SSCF2 plane and transition speeds produced by a native female speaker of Vietnamese.

Nous avons proposé que les gestes acoustiques soient définis par les trajectoires de SSF et la vitesse de SSCF selon les paramètres suivants :

- toutes les trajectoires SSCF sont des lignes assez droites ;
- la trajectoire de vitesse de la SSCF de chaque transition as une courbe elliptique plus ou moins marquée

Nous avons aussi proposé que les principaux paramètres dynamiques soient la direction de transition de chaque SSCF: en d'autres termes, ils peuvent être modélisés par les angles des évolutions pseudo-rectilignes des paramètres SSCF sur le plan fréquentiel (en supposant que toutes les trajectoires SSCF sont plus ou moins de lignes droites). Les angles SSCF peuvent être définis comme des angles entre les points de départ et les points de fin des trajectoires de transition dans le plan SSCF correspondant.

Par exemple sur le plan SSCF1/ SSCF2 de la Figure 4-18, les Angles SSCF12 sont des angles entre les points de départ et de fin de V1V2, trajectoires de transition dans le plan SSCF1/SSCF2, et sont signés par le paramètre α.

Sur chaque plan $SSCF_i/SSCF_{i+1}$, la formule de l'angle SSCF(i) (i + 1) est définie comme suit:

$$\text{angle(i)(i+1)} = \text{atan}\left(\frac{\Delta SSCF_{i+1}}{\Delta SSCF_i}\right) (^o) \quad \textbf{(4-2)}$$

Où:

$\Delta SSCF_{i+1}$ est la différence entre $SSCF_{i+1}$ à la fin et au début de la transition.

$\Delta SSCF_i$ est la différence entre $SSCF_i$, à la fin et au début de la transition.

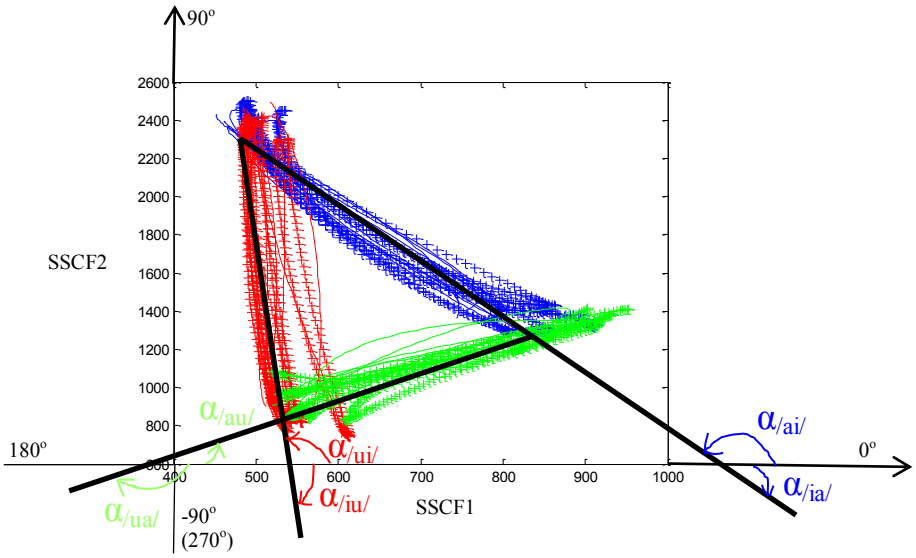


Figure 4-18: SSCF Angles12 in SSCF1/SSCF2 plane.

Analyse des angles SSCF pendant les transitions voyelle-voyelle pour la langue vietnamienne

Stimuli Vietnamien: Afin d'étudier les transitions vietnamiennes voyelle-à-voyelle, un petit corpus de transitions vietnamiennes V1V2 a été construit par huit locuteurs vietnamiens natifs (4 hommes et 4 femmes) qui sont nés et vivent dans le nord du Vietnam, avec des âges compris entre 25 et 32 ans. Les aspects de quatorze transitions / ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu / ont été étudiés. Chaque séquence V1V2 a été insérée dans une phrase porteuse: "Nói V1V2 ba lần" [noj^5 V1V2 ba^1 lɤn^2] ("Parler V1V2 trois fois»).

Chaque séquence V1V2 a été produite dix fois à des vitesses de parole différentes (5 fois au taux normal et 5 fois à vitesse rapide). Il y avait 1.120 phrases d'enregistrement pour huit locuteurs vietnamiens.

L'orateur a été chargé de lire les séquences V1V2 de telle sorte que le passage de la voyelle V1 à la voyelle V2 est continu dans le domaine temporel. Le processus d'enregistrement a été contrôlé par un

logiciel PC qui présente de façon aléatoire la succession des items à enregistrer. En cas de mauvaise prononciation ou d'hésitation, l'orateur doit prononcer à nouveau l'élément.

Cet enregistrement a eu lieu dans le studio calme de l'Institut Mica, à Hanoi, Vietnam. Le corpus a été acquis avec un système d'enregistrement pour une fréquence d'échantillonnage de 16.000Hz et de 16 bits par échantillon, et enregistré au format .wav.

L'orateur a été invité à lire le corpus à vitesse normale. Une petite pause a été faite entre la lecture des différentes phrases. Après avoir terminé le corpus à vitesse normale, le locuteur a été invité à lire le corpus à vitesse rapide. Ce processus se passe répétitivement dix fois au cours de l'enregistrement.

Méthode d'analyse

Afin de souligner le rôle de la discrimination des angles SSCF d'une séquence V1V2, pour chaque séquence V1V2, trois Angles SSCF ont été calculés comme suit:

- SSCF Angle12 est l'angle de V1V2 trajectoire dans le plan SSCF1/SSCF2;
- SSCF Angle23 est l'angle de V1V2 trajectoire dans le plan SSCF2/SSCF3;
- SSCF Angle34 est l'angle de V1V2 trajectoire dans le plan SSCF3/SSCF4;

Ensuite, ces trois angles ont été comparés en fonction des deux aspects suivants:

- le premier cas est la comparaison des angles SSCF entre les différents éléments pour chaque locuteur ;
- le second cas est la comparaison des angles SSCF au même point chez les hommes et les femmes.

Dans chaque cas, on compare les Angles SSCF entre le taux normal de parole et la vitesse rapide pour un même item et un même locuteur.

Résultats

Cas 1 : comparaisons des angles SSCF entre les différentes transitions pour chaque locuteur

Pour chaque locuteur, l'angle SSCF est plus ou moins variable en fonction à la fois de la vitesse d'élocution normale ou rapide pour chaque transition V1V2.

Pour chaque locuteur, les quatorze transitions peuvent être complètement séparées mutuellement dans le plan 3D des angles SSCF

Les figures 4-21 et 4-22, 4-23 et 4-24, 4-25 et 4-26 présentent l'écart moyen et le valeurs angulaires des paramètres SSCF Angle12, SSCF Angle23 et SSCF Angle34 pour les quatorze transitions /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/ produites respectivement par un homme vietnamien natif M1 et une femme vietnamienne F1 native.

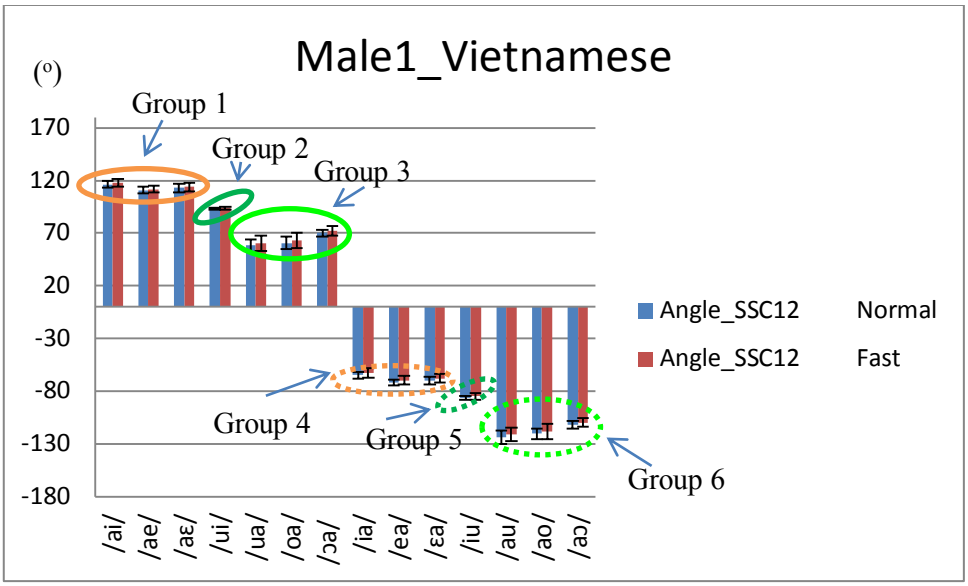


Figure 4-21: The average value and standard deviation of SSCF Angle12 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/produced by one Vietnamese male (M1) at both normal and fast rate.

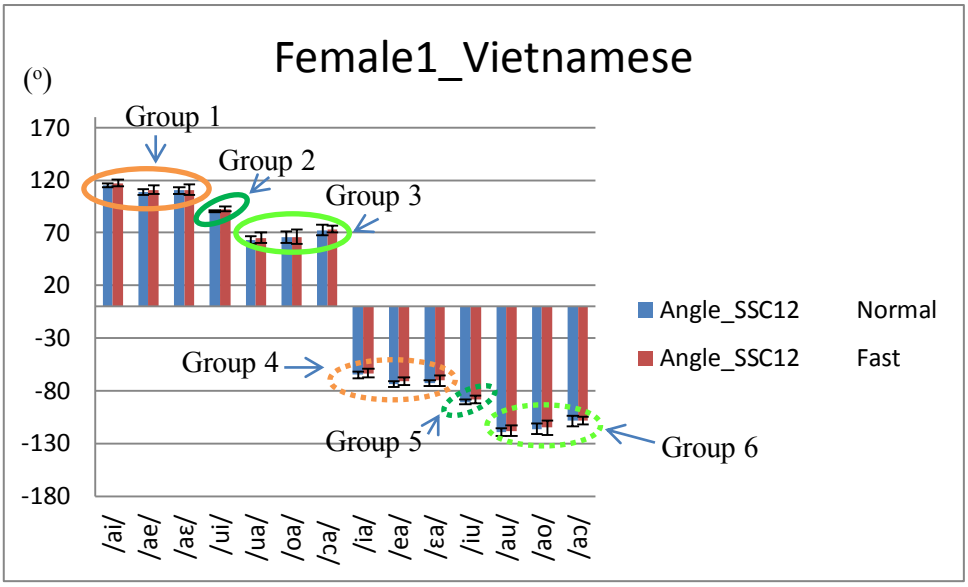


Figure 4-22: The average value and standard deviation of SSCF Angle12 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/produced by one Vietnamese female (F1) at both normal and fast rate.

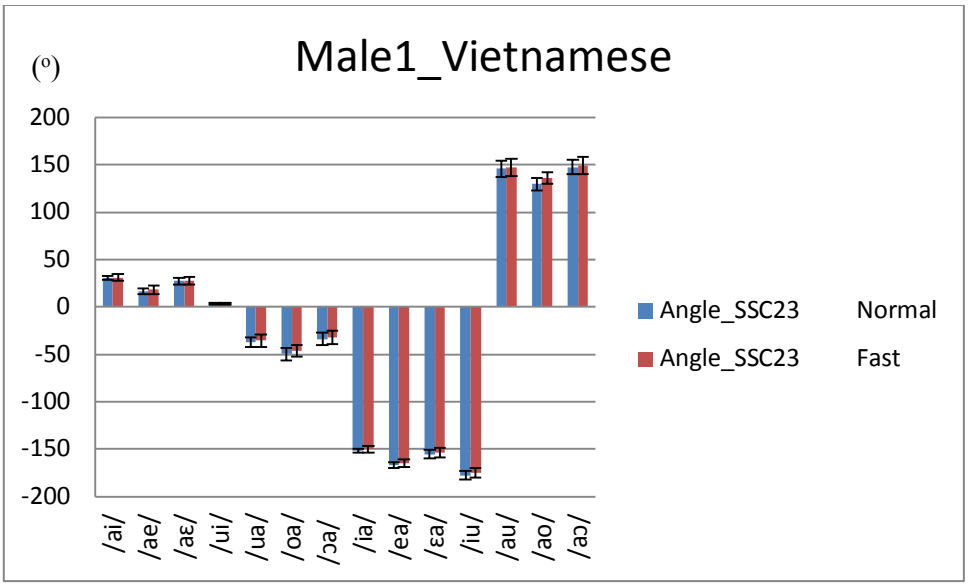


Figure 4-23: The average value and standard deviation of SSCF Angle23 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/produced by one Vietnamese female (M1) at both normal and fast rate.

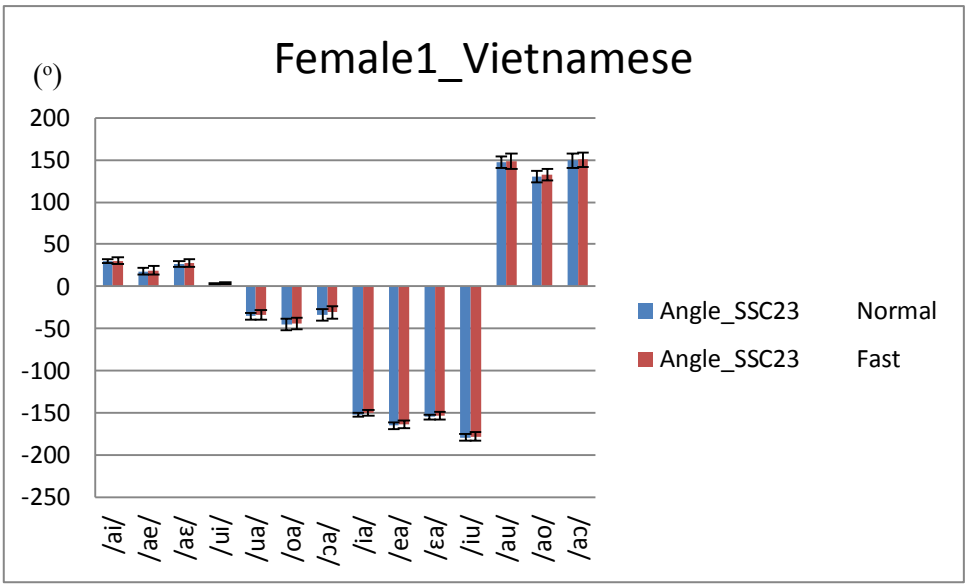


Figure 4-24: The average value and standard deviation of SSCF Angle23 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/produced by one Vietnamese female (F1) at both normal and fast rate.

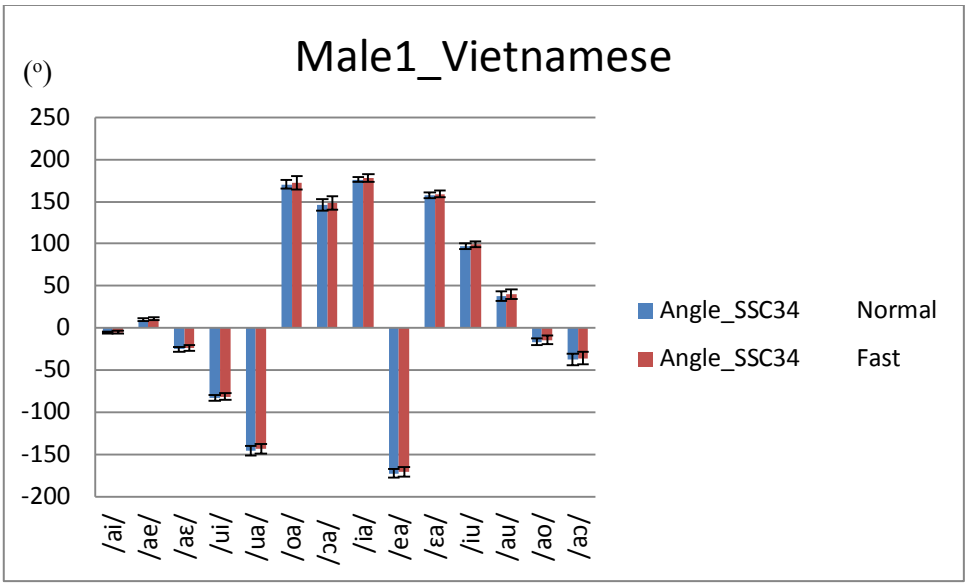


Figure 4-25: The average value and standard deviation of SSCF Angle34 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/produced by one Vietnamese female (M1) at both normal and fast rate.

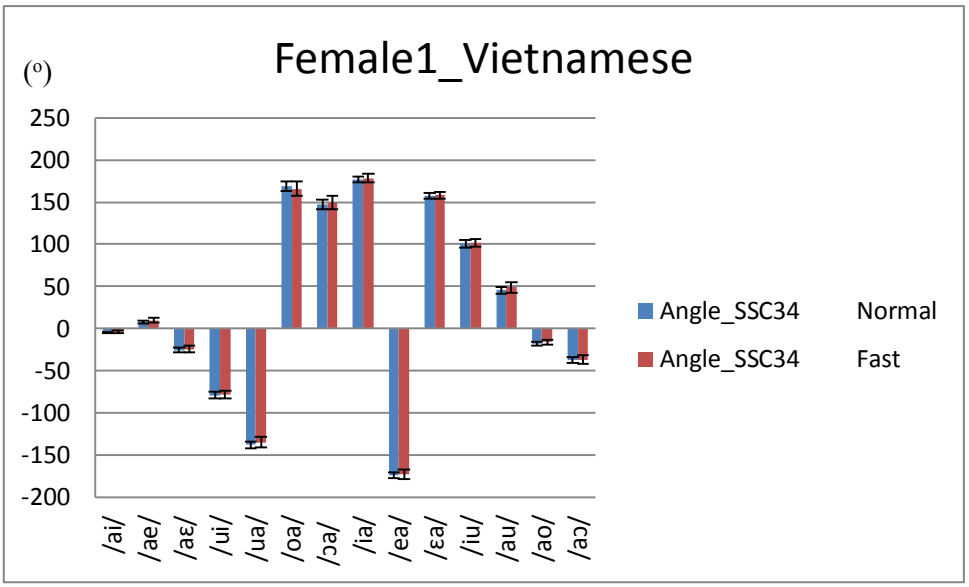


Figure 4-96: The average value and standard deviation of SSCF Angle34 of /ai, aɛ, ae, ua, ɔa, oa, ui, ia, ɛa, ea, au, aɔ, ao, iu/produced by one Vietnamese female (F1) at both normal and fast rate.

Cas 2: Comparaison des angles SSCF entre les hommes et les femmes

Pour les mêmes transitions V1V2, chaque Angle12 SSCF, Angle23 SSCF ou Angle34 SSCF présente plus ou moins la même valeur pour les huit locuteurs vietnamiens natifs étudiés (constitués de 4 hommes et 4 femmes). En d'autres termes, chaque angle SSCF ne dépend pas des différents locuteurs

Les figures 4-27, 4-28 et 4-29 4-27 présentent les résultats moyens et leur écart-type pour les paramètres Angle12 SSCF, Angle23 SSCF, et Angle34 SSCF d'un exemple de transition /ai/ produite par huit locuteurs vietnamiens (quatre hommes et quatre femmes).

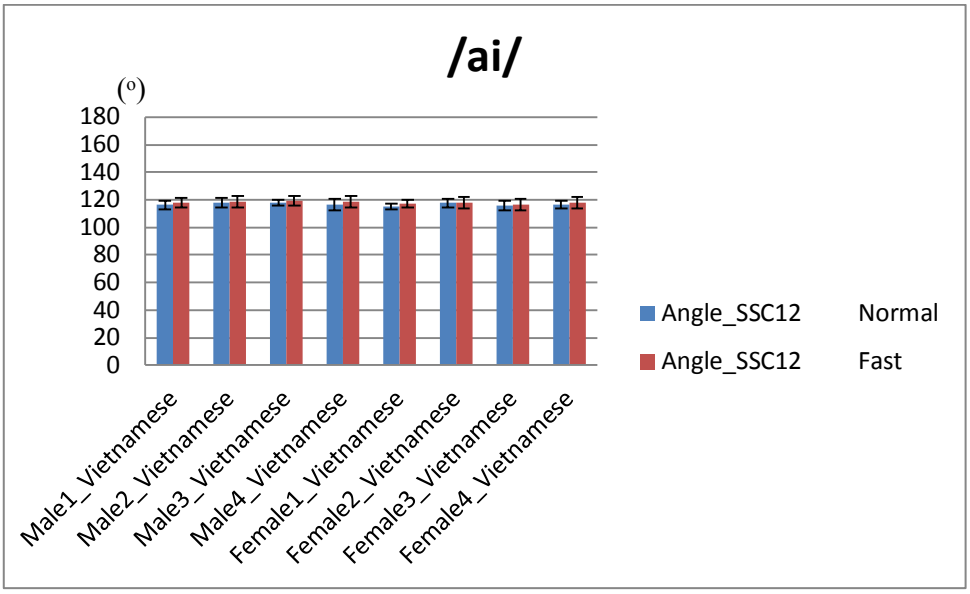


Figure 4-107: The average value and standard deviation of SSCF Angle12 of /ai/produced by 8 Vietnamese subjects (4 males + 4 females) at both normal and fast rate.

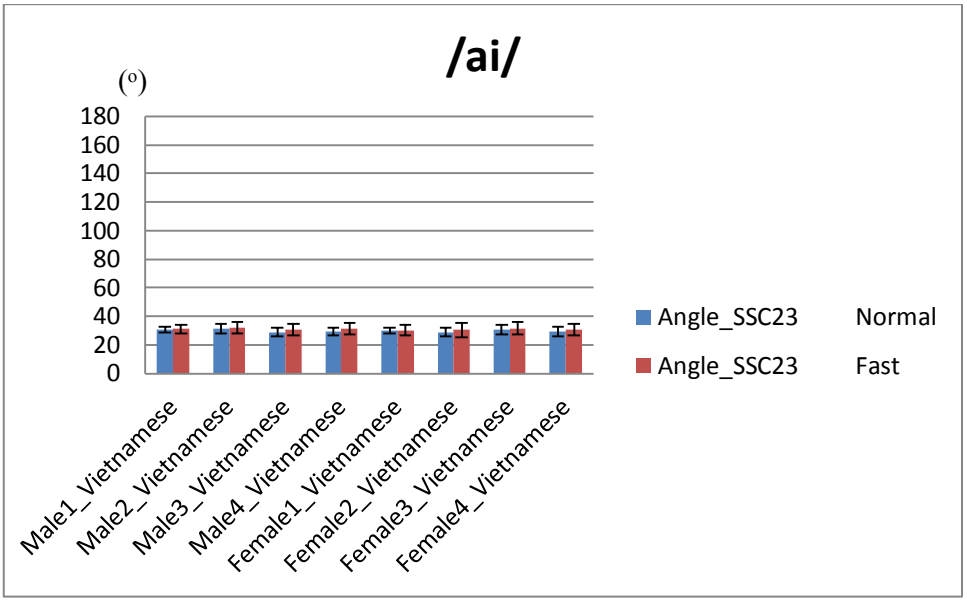


Figure 4-118: The average value and standard deviation of SSCF Angle23 of /ai/produced by 8 Vietnamese subjects (4 males + 4 females) at both normal and fast rate.

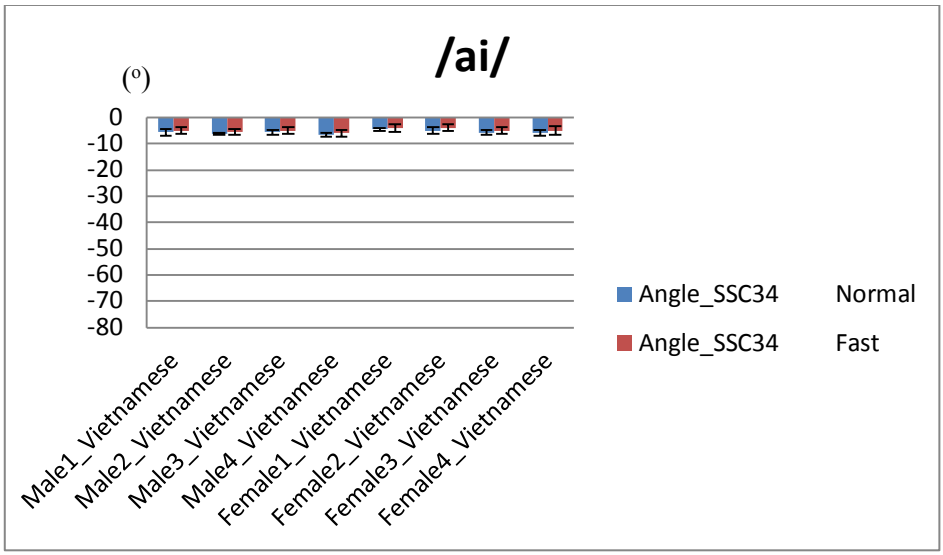


Figure 4-29: The average value and standard deviation of SSCF Angle34 of /ai/produced by 8 Vietnamese subjects (4 males + 4 females) at both normal and fast rate.

Des résultats similaires sont également obtenus pour les transitions en français /ai, ia, au, ua, ui, iu/.

Les résultats d'analyse sur les paramètres Angles SSCF ont également montré que chaque Angle SSCF sur la même séquence de transition V1V2 présente plus ou moins de la même valeur pour les hommes et les femmes. Et ils sont aussi assez invariants pour le débit de parole (taux de la parole normale et rapide) pour chaque locuteur.

Sur cette base, nous proposons de vérifier la possibilité d'utiliser ces propriétés d'indépendance du locuteur des Angles SSCF pour un système de reconnaissance vocale automatique vietnamien. Cette tâche sera présentée dans le **Chapitre 5**.

Les résultats attendus concernent fondamentalement le fait que le système de reconnaissance vocale RAP construit sur cette base serait beaucoup moins dépendant du locuteur: un petit nombre de locuteurs serait nécessaire pour arriver à un échantillon suffisant et adéquat. Par conséquent, ceci diminuerait considérablement la quantité de données qui est nécessaire pour la formation du système RAP.

A cet effet, nous avons décidé d'utiliser un système de reconnaissance automatique de la parole simple et très classique pour la langue vietnamienne qui a été développé à l'Institut de recherche MICA

Le but n'est pas ici d'obtenir un excellent système de reconnaissance, mais de comparer nos paramètres d'angles SSCF proposés avec les paramètres MFCC classiquement utilisés dans la reconnaissance vocale.

Nous avons divisé nos tests en deux parties:

Tout d'abord, nous tenons à vérifier que les angles SSCF sont utilisables comme entrée indépendante du système RAP. Dans ce test, nous utilisons un corpus équilibré en comparaison avec un MFCC classique.

Deuxièmement, nous souhaitons vérifier que le système RAP utilisant les paramètres Angles SSCF proposés est moins dépendant du locuteur. Dans ce test, nous utilisons un corpus déséquilibré : un modèle formé de voix masculines a été appliquée à la reconnaissance des voix féminines; et vice versa.

Cas 1 : Les angles SSCF sont utilisés comme entrée indépendante du RAP

Corpus

Le type de parole choisi est un long paragraphe commun. Le corpus comprend 12 locuteurs natifs vietnamiens du Nord et il est divisé pour assurer l'équilibre entre les voix masculines et féminines pour tous les deux entrainement et décodage. Notre approche est habituellement utilisée pour effectuer des tests sur les systèmes RAP afin de réduire la différence de caractéristiques de parole entre les voix masculines et féminines. Suite à cette approche, le corpus est divisé comme suit :

- base de données d'entraînement : 3,9 heures de parole de 8 locuteurs vietnamiens (4 hommes et 4 femmes) ;
- base de données du décodage : 0,9 heures de parole de 4 locuteurs vietnamiens (2 hommes et 2 femmes).

Méthode

Nous utilisons l'outil Sphinx pour effectuer le test RAP sur le corpus vietnamien décrit ci-dessus. Rappelons que notre système RAP utilise seulement un modèle acoustique (sans modèle de langage). Comme le vietnamien est une langue tonale, nous utilisons un RAP fonctionnant avec des phonèmes tonaux vietnamiens.

Résultats

Les résultats pour la reconnaissance de la parole vietnamienne de MFCCs, des angles SSCF et leurs dérivées premières et secondes, sont présentés dans les tableaux 5-2 et 5-3.

Pour les deux systèmes ASR, si l'on utilise un vecteur de paramètres simples, les scores de SER sont similaires: SER = 29,050% avec des paramètres MFCC et SET = 28,369% avec des paramètres d'Angles SSCF . Toutefois, dans ce dernier cas, le nombre de paramètres pris en compte est divisé par deux (taille du vecteur SSCF Angles = 6, par rapport à la taille du vecteur MFCC = 13). On peut considérer que l'on obtient des résultats similaires, même si la caractérisation du signal dans le domaine fréquentiel contient moins d'informations utiles. Cela constitue un avantage en termes de complexité et de temps de calcul.

Table 5-2: Syllable error rate (%) in Vietnamese ASR using MFCC and their derivation with the balance between male and female voices in both training and testing sets.

Types of parameters	SER (%)	Dimension of parameter vector
MFCC	29.050	13
Δ + ΔΔ	11.732	26
MFCC + Δ + ΔΔ	7.952	39

Table 5-3: Syllable error rate (%) in Vietnamese ASR using SSCF Angles and their derivationwith the balance between male and female voices in both training and testing sets.

Types of parameters	SER (%)	Dimension of parameter vector
SSCF Angle	28.369	6
SSCF Angle + Δ	12.212	12

Si nous ajoutons la dérivée première (Δ) des paramètres d'Angle SSCF, le score SER est alors similaire aux cas où seules les caractéristiques de dérivées (dynamique) des signaux (Δ et Δ Δ) sont calculées à partir de paramètres MFCCs (respectivement 12.212% et 11,732%). Ce résultat confirme notre hypothèse selon laquelle notre façon de caractériser les caractéristiques dynamiques produit des résultats similaires par rapport à l'approche classique.

En première conclusion, nous pouvons écrire que ces résultats montrent que les paramètres SSCF Angles peuvent être utilisés comme entrée directe du système de RAP.

Cas 2 : le système RAP utilisant les paramètres proposés d'Angles SSCF est moins dépendant du locuteur

Corpus

Le type de parole choisi est composé de longs paragraphes communs pour des locuteurs natifs vietnamiens du Nord. Ce corpus est divisé non équitablement entre les voix masculines et féminines dans la base de données d'entraînement et de test :

Pour le premier test, la base de données d'entraînement comprend 2,5 heures de parole produites par 6 locuteurs hommes vietnamiens. La base de données du décodage consiste en 0,5 heures de parole de 2 locutrices femmes vietnamiennes.

Pour le deuxième test, nous avons fait le contraire : la base de données d'entraînement comprend 2,4 heures de parole produites par 6 locutrices femmes vietnamiennes La base de données du décodage consiste en 0,45 heures de parole de locuteurs hommes vietnamiens.

Les deux versions du système ASR utilisent uniquement un modèle acoustique.

Résultats

Premier test asymétrique (entraînement avec les hommes et test avec les femmes). Les résultats RAP vietnamiens sur ce test sont présentés dans les tableaux 5-4 et 5-5, comme suit:

Table 5-4: Syllable error rate (%) in Vietnamese ASR using MFCC and their derivations with the male training and female testing.

Types of parameters	SER (%)	Dimension of parameter vector
MFCC	74.505	13
Δ + ΔΔ	42.162	26
MFCC + Δ + ΔΔ	41.223	39

Table 5-5: Syllable error rate (%) in Vietnamese ASR using SSCF Angles and their derivations with the male training and female testing.

Types of parameters	SER (%)	Dimension of parameter vector
SSCF Angle	48.907	6
SSCF Angle + Δ	25.444	12

Une comparaison entre les meilleurs résultats obtenus par le système RAP en utilisant des paramètres MFCC (MFCC + Δ + = ΔΔ→SER = 41,223%, dans le tableau 5-4) et ceux obtenus par la version RAP en utilisant les paramètres d'angle SSFC montre que le système utilisant ces derniers paramètres est bien meilleur (taux d'erreur presque deux fois plus petit: SER = 25,444%). En outre, le fait que dans le premier cas, le vecteur d'entrée contient 39 paramètres alors que dans le second cas, il ne contient que 12 paramètres, démontre que le second système qui utilise les angles SSCF + Δ est meilleur.

Cette observation nous suggère que les caractéristiques vocales acoustiques dynamiques (caractérisées par nos paramètres angles SSCF) fonctionnent beaucoup mieux que les caractéristiques acoustiques pseudo-statiques habituellement utilisées (MFCC).

Des résultats similaires ont été obtenus avec le deuxième test asymétrique (entraînement avec les femmes et test avec les hommes). Les résultats RAP vietnamiens sur ce test sont présentés dans les tableaux 5-6 et 5-7, comme suit:

Table 5-6: Syllable error rate (%) in Vietnamese ASR using MFCC and their derivations with the female training and male testing.

Types of parameters	SER (%)	Dimension of parameter vector
MFCC	92.071	13
Δ + ΔΔ	54.179	26
MFCC + Δ + ΔΔ	50.536	39

Table 5-7: Syllable error rate (%) in Vietnamese ASR using SSCF Angles and their derivations with the male training and female testing.

Types of parameters	SER (%)	Dimension of parameter vector
SSCF Angle	42.875	6
SSCF Angle + Δ	31.911	12

Par conséquent, nous avons démontré qu'il est possible d'utiliser les paramètres d'angle SSCF en tant que vecteur d'entrée du système de reconnaissance de la parole, et que les ensembles de résultats obtenus sont très semblables à ceux d'un système de reconnaissance classique.

Nous avons réalisé des tests de reconnaissance avec des bases de données asymétriques en fonction des genres des locuteurs. Les résultats ont montré que notre système ASR, en utilisant les angles SSCF, est beaucoup moins dépendant des locuteurs que ceux qui utilisent les MFCCs. Cela confirme notre hypothèse que la caractérisation des gestes acoustiques dynamiques peut être un grand avantage pour la reconnaissance automatique de la parole, car il permet la conception des systèmes de reconnaissance vocale qui sont intrinsèquement indépendant des locuteurs.

Ces premiers tests, même s'ils sont très intéressants, ne suffisent pas. Ils ont montré que nous devions améliorer notre système de reconnaissance. Les premières orientations possibles pour les améliorations sont les suivantes: (i) un meilleur calcul des angles, plus proche de la théorie, afin d'obtenir un vecteur d'entrée plus représentatif du geste acoustique; (ii) une meilleure caractérisation de ces gestes acoustiques lors de la production des consonnes; (iii) une meilleure prise en compte de la vitesse de transition, non seulement dans le calcul de la dérivée de l'angle, mais aussi dans le calcul d'un angle spécifique de "vitesse", mesuré directement sur les transitions de vitesse.

Enfin, nous sommes conscients que nous devons faire nos tests avec un corpus d'entrainement beaucoup plus large enregistrés par plus de locuteurs pour obtenir une meilleure représentation de la langue. Les résultats seront plus fiables et permettront par la suite de meilleures comparaisons.