MODELING OF PLOSIVE TO VOWEL TRANSITIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ALİCAN BEKÖZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

AUGUST 2007

Approval of the thesis:

MODELING OF PLOSIVE TO VOWEL TRANSITIONS

submitted by **ALİCAN BEKÖZ** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İsmet Erkmen _____
Head of Department, **Electrical and Electronics Engineering**

Prof. Dr. Mübeccel Demirekler _____
Supervisor, **Electrical and Electronics Engineering Dept., METU**

Examining Committee Members:

Assoc. Prof. Dr. Tolga Çiloğlu _____
Electrical and Electronics Engineering Dept., METU

Prof. Dr. Mübeccel Demirekler _____
Electrical and Electronics Engineering Dept., METU

Asst. Prof. Dr. Afşar Saranlı _____
Electrical and Electronics Engineering Dept., METU

Assoc. Prof. Dr. Gökhan İlk _____
Head of Electrical and Electronics Engineering Dept., AU

Dr. Özgül Salor _____
TÜBİTAK

Date: (28.08.2007)

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Alican BEKÖZ

Signature :

ABSTRACT

MODELING OF PLOSIVE TO VOWEL TRANSITIONS

BEKÖZ, Alican
M.Sc., Department of Electrical and Electronics Engineering
Supervisor: Prof. Dr. Mübeccel Demirekler

August 2007,145 pages

This thesis presents a study concerning stop consonant to vowel transitions which are modeled making use of acoustic tube model.

Characteristics of the stop consonant to vowel transitions are tried to be obtained first. Therefore several transitions including fricative to vowel transitions are examined based on spectral and time related properties. In addition to these studies, x-ray snapshots, lip videos and also experiments including subjects are used to intensify the characterization, from the production and the perception side of views. As results of these studies the plosive to vowel transitions are observed to be uttered by exponential vocal tract movements and the perception mechanism is observed to be highly related with exponential spectral changes.

A model, based on the acoustic tube model, is tried to be established using the knowledge and the experience gained during characterization therefore proposed model involves the vocal tract parameters observed in characterization part.

Finally, plosive to vowel transitions including three types of plosives (alveolar, labial and velar) are synthesized by the proposed model. The formants of the synthesized sounds are compared with the formants of the natural sounds.

Also the intelligibility tests of these sounds are done. Performance evaluation tests show the proposed model's performance to be satisfactory.

Keywords : **plosive, stop consonant, model**

ÖZ

PATLAMALI HARFLERDEN ÜNLÜ HARFLERE GEÇİŞİN MODELLENMESİ

BEKÖZ, Alican
Yüksek Lisans, Elektrik Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Mübeccel Demirekler

Ağustos 2007, 145 sayfa

Bu tez akustik tüp modelinden yararlanılarak patlamalı harflerden ünlü harflere geçişin modellenmesini sunar.

Öncelikle patlamalı harf ünlü harf geçişlerinin karakteristikleri tanımlanmaya çalışılmıştır. Bu amaçla frikatif harf ünlü harf geçişleri de dahil olmak üzere bir çok geçiş spektral ve zaman bazlı özellikler yönünden incelenmiştir. Bu çalışmalara ek olarak incelemeleri boğumlama ve algılama yönlerinden kuvvetlendirmek amacıyla X-ray görüntüleri, dudak videoları ve denek kullanılan deneylerden yararlanılmıştır. Bu çalışmaların sonucunda patlamalı ses ünlü ses geçişlerinin üstsel ses yolu hareketlerinden oluşturulduğu ve algılama mekanizmasının üstsel spektrum değişiklikleriyle yakından ilgisi olduğu gözlemlenmiştir.

Elde edilen bilgi ve deneyimlerle akustik tüp modellemesine dayalı bir model oluşturulmaya çalışılmıştır. Bu nedenle kurulan model tanımlama kısmında gözlemlenen ses yolu parametrelerini içermektedir.

Son olarak üç çeşit patlamalı harflerden (dudaksıl, dişyuvasıl ve damaksıl) ünlü harflere geçişler öne sürülen model kullanılarak sentezlenmiştir. Sentezlenen geçişlerin formant frekansları doğal geçişlerin formant frekanslarıyla karşılaştırılmıştır. Ek olarak bu seslerin anlaşılabilirlik testleri de yapılmıştır.

Performans hesaplama testleri modelin performansını tatminkar olarak göstermiştir.

Anahtar Kelimeler : **patlamalı harf, süreksiz sessiz harf, model**

To my family and Gökçe

ACKNOWLEDGMENTS

The author wishes to express his deepest gratitude to his supervisor Prof. Dr. Mübeccel Demirekler for her invaluable guidance, advice and support throughout the research.

The author would also like to thank Dr. Özgül SALOR and Mr. Yücel ÖZBEK for their beneficial help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Literature Review and Thesis Outline

Concatenated acoustic tube modeling is a powerful method which acts as the basis of many speech production models. Derivation of that model is based on the pair of differential equations proposed by Beranek [33] (making use of Newton physics), which describes the pressure and the volume velocity of the propagating sound waves in the concatenated tubes. Making use of the differential equations, Kelly and Lochbaum [2] were the first people who used the reflection coefficient definition that describes the reflected and the propagated waves at each junction. They also proposed the signal flow graph representation in which the vocal tract can be modeled as a series of concatenated tubes that describes the sound propagation with the reflection coefficients at each junction. After all these studies, a complete model is constructed by adding excitation and radiation properties for speech synthesis.

Acoustic tube modeling simply uses the quantized vocal tract area functions in the order of few tens. Hence, to obtain a linear model based on reflection coefficients, vocal tract areas should be obtained. In the literature various vocal tract visualization techniques are developed. First studies started with the use of X-ray. Öhman and Stevens [8] were among the first researchers who used X-ray films for vocal tract visualization in 1960s. The method was simple; subjects were given texts to utter and simultaneously the X-ray videos were recorded. In addition to the X-ray method, other methods like electromyography or a velar trace are also used, however these methods including X-ray are either hazardous to human

health or interfere the articulation progress because of the sensors located in the vocal tract.

A more accurate solution to obtain the vocal tract area functions without hazarding the human health or interfering the utterance became possible with the use of magnetic resonance imaging known as MRI in 1990s. Baer et al. [16], Perrier et al. [11] and Mohammed [6] were examples of the researchers visualizing the vocal tract for vowels making use of MRI. In addition to the vowels, plosives, fricatives and nasals are also visualized by Shrikant et al. [17] and Story et al. [25] in a stationary mood with a virtual vowel target. However, the MRI technique was useless for continuous transitions, in addition, the method was also much time consuming. Therefore some techniques are developed in order to estimate the vocal tract area functions from cross sectional sagittal distances to make use continuous visualization techniques. Studies of Baer et al. [16], Maeda [26], Mermelstein [7] and Perrier et al. [11] also include some techniques to segment the vocal tract properly for quantization of the continuous vocal tract. As a general method, alpha beta model which describes the vocal tract areas in terms of sagittal distances for specified regions of the vocal tract is proposed. However these methods are limited in their use since they are subject dependent thus the results are not very reliable.

On the other hand, use of all these techniques and much powerful computer systems improved acoustical analysis of speech. Fant [27], Flanagan [28], Rabiner and Schafer [1] were the leading speech related researchers who worked with acoustical theory of speech production and properties of the speech. At the same time, specified studies about fricatives, vowels, nasals and plosives are also done. As the main focus of this study, the plosives are first studied by Fischer and Jorgensen [24] in 1950s. In this study acoustical properties of stop consonants, which differentiate them from other sounds, are analyzed with experiments. In 1952 Cooper et al. [14] investigated the perceptional properties of plosives by making controlled studies on various subjects. According to them, the perception of plosive to vowel transitions depends on the formant differences between the beginning of the transition and steady state vowel. After these studies Halle et al. [23] also investigated characteristics of stop consonants. In 1970s Stevens [20] (and Klatt) [21] [22] and Victor [9] performed acoustical studies about plosives examining burst, aspiration and articulation characteristics of these sounds. In

recent years, Suchato [10] and Jackson [4] [5] made statistical studies about characterization of these sounds. As the main results of these studies, the dominancy of the vowels in plosive to vowel transitions is figured out. Also plosive to vowel transitions are agreed to include sudden formant changes that are uttered by oral constrictions. In addition, the tracks are observed to be exponential by Jackson's studies.

In this work, a plosive to vowel transition is tried to be realized using Kelly Lochbaum [2] acoustic tube model in a state space representation. Any other parameters such as excitation and radiation are used as the general ones that exist in the literature. First of all, general information is obtained about plosives by related studies and making use of literature. In addition to the literature, statistical studies about the properties of plosives have been done. The dominancy of following and preceding vowel, the character of the formant tracks and the effect of formant differences on the perception of these sounds are found to be in an agreement with the existing literature. Moreover, for a precise analysis of these sounds, X-ray snapshots [8] of these transitions including three types of the plosives (labial, alveolar and velar) are investigated. Sagittal distances are measured by using digital pixel data of each snapshot. As a result of this study, the sagittal distance changes are observed to be similar to exponential functions with time constants within a proper range. In addition to the X-ray images, fast frame rate lip videos are captured for better analysis of alveolar and labial regions. This study also showed similar results with X-ray investigations. Moreover, the time constants of the alveolar and labial video snapshots in both transitions (vowel to plosive, plosive to vowel) are found to be similar with the formant transitions of the same utterances.

Having common properties with plosives, fricatives are also examined. Analysis of formants of fricatives showed great similarity to formant trajectories of the plosives of the same type. The only difference is at the region where partial closure occurs. Manually truncation of these regions of the fricatives using the time waveform resulted a perception of the plosive of the same kind (as an example, truncated labial fricative is percept as labial plosive). Such an interesting perception of the plosives is due to the similarity of the vocal tract movements for both cases. Vocal tract movement similarity is also inspected in X-ray snapshots. In addition to the X-ray snapshots, lips video examination has also been done for

fricatives. Exponential change in area function is observed, starting from the moment that it is percept as the plosive counterpart. Also this interval is observed to have similar time constant compared to the same type of plosive utterance. As a result, plosives are observed to be the inhibited versions of the fricatives by total oral constrictions.

The model used in all of the above mentioned studies is obtained from etiher X-ray videos of the vocal tract or lip videos of a day camera. For a derivation of the model, area functions are needed instead of the sagittal distances. Therefore normalized areas for each tube are obtained according to the alpha beta model with a proper range commonly used in the literature [11] [13] [15]. Estimated areas are also observed to be exponential like functions as expected and theoretically illustrated. In addition to the area functions, reflection coefficients are also calculated making use of estimated areas. The reflection coefficients are observed to be almost stationary except the ones that correspond to the constriction region during the transition. This is also confirmed by MRI data [25] of plosives in the context of different vowels. On the other hand, the coefficients at the constriction region tend to change exponentially between the plosive and the vowel configurations. The vocal tract resonances obtained by changing only two reflection coefficients (one for infinite decrease and other for infinite increase in consequent areas) of modified vowel to the original vowel configuration yielded similar formant tracks compared to the natural ones and even intelligibility tests give satisfactory results. For further analysis, all reflection coefficients are changed from plosive's to vowel's reflection coefficients. The vocal tract resonances of the synthetic utterances showed great resemblance to the original tracks, also recognitions rates are observed to be satisfactory.

As a last study, plosive to vowel transitions are synthesized according to the exponential area model by making use of MRI data [25] (MRI data is used to calculate the plosive areas if data is available, otherwise they are estimated making use of the vowel data). Synthesized transitions are observed to have similar formant tracks with the natural ones. Also they are distinguished by the listeners with a high percentage.

This thesis is organized as follows, in the second chapter state space representation of the entire model and basic acoustic properties of stop consonants are given. Continuing with the third chapter, several examinations of

different aspects are done for plosive to vowel characterization. This chapter includes spectral, time domain and visual (X-ray, day camera) examinations. Then in the fourth chapter a model is tried to be proposed according to the experience gained in Chapter 3. Results follow this chapter for a performance evaluation and finally comments and conclusion is presented.

CHAPTER 2

THEORETICAL BACKGROUND

2.1. State Space Representation of the Entire Model

In our study, Kelly and Lochbaum [2] speech synthesis model is used as mentioned before. This model is explained in Appendix in details.

In this study we have used a state space representation. N[th] order concatenated tube vocal tract model with states x_1 to x_n is shown in Figure A-13 (See Appendix A.1.3.2). For every state, following set of equations can be obtained by using general signal flow graph properties:

$$
\begin{aligned}
&x_1(n+1) = -r_G r_1 x_1(n) - r_G(1-r_1)\, r_2 x_2(n) - .... - r_G(1-r_1)(1-r_2)...(1-r_{n-1}) x_N(n) + (1+r_G)/2 u_G(n);\\
&x_2(n+1) = (1+r_1) x_1(n) - r_1 r_2 x_2(n) - .... - r_1(1-r_2)..(1-r_{n-1})\, r_l x_N(n);\\
&\quad .\qquad .\\
&\quad .\qquad .\\
&x_N(n+1) = (1+r_{N-1})\, x_{N-1}(n) - r_{N-1} r_L x_N(n);
\end{aligned}
\tag{2.1}
$$

Letting:

$$
\hat{X}_k = \begin{bmatrix} x_1(k) & x_2(k) & x_3(k) & \cdots & x_N(k) \end{bmatrix}' \tag{2.2}
$$

We obtain the following state space representation:

$$
\begin{aligned}
\hat{X}_{k+1} &= A\hat{X}_k + B u_G\\
\hat{Y}_k &= C\hat{X}_k
\end{aligned}
\tag{2.3}
$$

Where:

$$A=\begin{bmatrix} -r_G r_1 & r_G(1-r_1)\ r_2 & r_G(1-r_1)(1-r_2)\,r_3 & \cdots & -r_G\ r_L\prod_{i=1}^{N-1}(1-r_i). \\ (1+r_1) & -r_1\ r_2 & -r_1\ (1-r_2)\ \ r_3 & \cdots & -r_1\ r_L\prod_{i=2}^{N-1}(1-r_i) \\ 0 & (1+r_2) & -r_2\ \ r_3 & \cdots & -r_2\ r_L\prod_{i=3}^{N-1}(1-r_i) \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & -r_{N-1} \end{bmatrix}$$

$$B=\begin{bmatrix}0.5(1+r_G) & 0 & \cdots & 0\end{bmatrix}^{\mathbf{T}} \qquad C=\begin{bmatrix}0 & \cdots & 0 & (1+r_L)\end{bmatrix} \tag{2.4}$$

In Equation 2.3, $\hat{Y}_k$ is the volume velocity at the lips end. To have the pressure at the lips end, the radiation should be added to the state space model. The radiation can be modeled as a FIR filter whose impulse response has the z transform (See Appendix A.1.4.) $R(z)=1-\alpha z^{-1}$ where, generally the pole of the filter is almost on the unit circle meaning that α is very close to 1. Cascading the radiation to the vocal tract model, following pressure equation is obtained.

$$\hat{P}_k=\hat{Y}_k-\alpha\hat{Y}_{k-1}\ \text{: output} \tag{2.5}$$

Let

$$\bar{\xi}_k=\begin{bmatrix}\hat{X}_k\ \hat{Y}_{K-1}\end{bmatrix}' \tag{2.6}$$

Rewriting matrix equations:

$$\bar{\xi}_{k+1}=\begin{bmatrix}A & 0\\ C & 0\end{bmatrix}\bar{\xi}_k+\begin{bmatrix}B\\0\end{bmatrix}u_G \qquad \hat{P}_k=\begin{bmatrix}C & -\alpha\end{bmatrix}\bar{\xi}_k \tag{2.7}$$

The state space representation has advantages over the general AR filter of the vocal tract. First of all, the effects of parameters over the filter are easier to figure out. Secondly, it is easier to synthesize speech corresponding to non-

stationary vocal tract and finally, the excitation can be inserted from any desired articulation state which may be helpful for analysis and synthesis purposes.

To sum up, state space representation is the matrix form of the vocal tract filter. The advantages of this representation over pole zero representation of the AR filter are its flexible and analyzable structure. Thus, this representation is used in this study for analysis and synthesis purposes.

2.2. Acoustic Properties of Stop Consonants

The aim of this part is to provide some background on utterance of stop consonants in a more morphological manner. The mathematical model of speech is given in the Appendix. The model consists of a source-filter approach in which all acoustic events are explained both in time domain and in frequency domain. In the following sections, the three types of stop consonants used commonly in many languages, labial, alveolar and velar stop consonants, are compared in terms of distinctive acoustic properties that are expected to be useful in discrimination of them. The characteristics of these plosives and the characteristics of factors involved in them are explained in a qualitative manner. The voiceless versions of these stops are also analyzed for the sake of completeness.

2.2.1. The Production of Stop Consonants

In Appendix, it is explained that, the speech production process can be realized by a source-filter model. The source corresponds to the mechanism that generates the air flow whereas the filter corresponds to the path that the air flows through. The origin of airflow is actually the lungs, whereas the path is formed by the trachea, larynx, pharynx, oral cavity and for some situations, nasal cavity (See Figure 2-1 [19]). In a simplest way the source and path operate as follows, the diaphragm pushes the ribcage to pump out the accumulated air in the lung afterwards trachea transmits the pressurized air leaving the lungs to the larynx, which is called the voice box of the articulatory system. The reason of this nickname that is embroidered to larynx is the presence of vocal folds within its structure. The function of the vocal folds is simply to process the air flow for voiced

utterances as known. This is achieved by simple vibrations of the vocal folds opening and closing with a frequency band typically 80 – 250 Hz depending on the genre and as a result, a waveform resembling to periodic pulse train as the one given in Appendix A.1.5. is obtained at the glottis output. Whereas, if the utterance is an unvoiced one the larynx gets narrowed and bypassed for unvoiced excitation. The generated excitation follows pharynx to the oral cavity then radiates from the mouth opening and the nostrils (if nasal coupling occurs) leaving the oral cavity. While the air flows, the shape of this path is dynamically controlled by the movement of various articulators along its length, such as glottis, velum, tongue, teeth and lips. The movements of these articulators combined with movement of the overall vocal tract path (chin and remaining articulator muscles), dynamically changes the parameters of the proposed vocal tract filter which are explained in the Appendix. The manipulation of these parameters changes the response of the filter, thus the articulatory system gains the capability of producing various sounds with different characteristics.

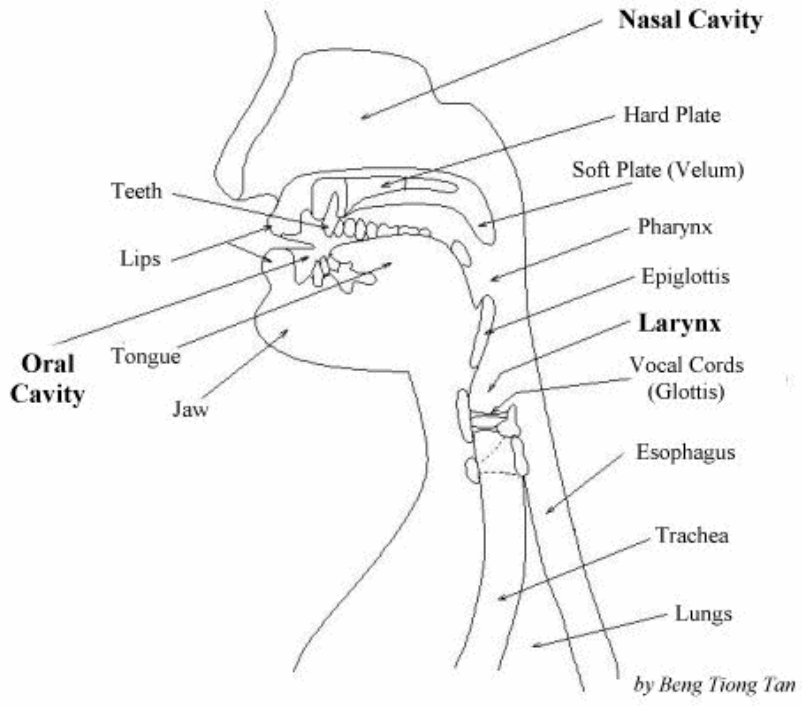

Figure 2-1 : Human Articulatory System.

One group of these sounds, the plosives, are produced with the sudden releases in the vocal tract. A human being utters a stop consonant by one primary articulator, an articulator in the oral cavity, forming a total closure of the vocal tract simultaneously preserving the pressure of the air in the lungs. The blockage causes the pressure of the air flow to increase at the point before the constriction. At this state, a silence occurs at the output of the oral cavity. An exception to this inhibition may occur for voiced plosives, in such a way that the inhibited source radiates from the soft skin before the constriction place instead of using the air path, producing a low frequency wave so called *voice bar*. The voice bar can simply be assumed as the pre-processed version of the excitation, that's why many similarities exist with the pure excitation. Leaving the details and concentrating on the process again, the vocal tract relaxes the constriction rapidly, causing the rushing of air through the just-released region. At this stage, noise is generated at the constriction due to the rapid moving of the air through the small opening [20]. This situation can be resembled to a flat ball or a tire generating a burst like noise at the cracking stage because of the pressurized air that rushes out. Thus, the airflow generated by the noise is called as the burst part of the stops. After the burst, if the stop consonant is instantly followed by a vowel, the vocal folds will start vibrating again for vowel utterance. The elapsed time between the release of the constriction and start of the glottal excitation is called as the voice onset time and will be called *VOT* from now on (The interval will be interpreted as VOT region). The VOT is in the order of few ten milliseconds which is typically longer for unvoiced stops. The reason behind this property can be related to the characteristics of the unvoiced plosives. Unlike voiced plosives, the unvoiced plosives are uttered by unvoiced excitations, therefore the excitation is kept in non vibrating mode because of the nature of these sounds resulting a longer interval. In addition to this, the requirement to release the turbulent noise that exists in the vocal tract, takes extra time to fulfill the function so called the *aspiration* .Sometimes the aspiration may also be called as the aspiration noise because of its noisy characteristics.

The above mentioned events can be observed in the spectrogram of the utterance /g ae g/ given in Figure 2-2 [10]. Total closure is formed by tongue body and the posterior hard palate in the region marked by (1). The closure inhibits the air flow that corresponds to the inhibition of high frequency components. At this

stage, only low frequency energy radiated from the soft vocal tract walls, resembles excitation like waveform mentioned before. The darkness at the spectrogram at the frequency range of 100-200 Hz indicates the existence of a low frequency periodic waveform supporting the claims. Next, in the region (2), the constriction is relaxed and sudden airflow rushes through the small opening causing the release burst. This section is actually the burst and the aspiration part and is very short because of the voiced excitation. As a last step, in region (3), the glottis starts vibrating again. Simultaneously the vocal tract shape is driven from the configuration at the closure to the configuration of the following vowel. During this interval the difference of two configurations causes a smooth transition between the plosive and the vowel. The details of this transition will be explained later but as a word, the continuity of the formants shows the importance of the succeeding vowel in order to understand the nature of plosive to vowel transitions.

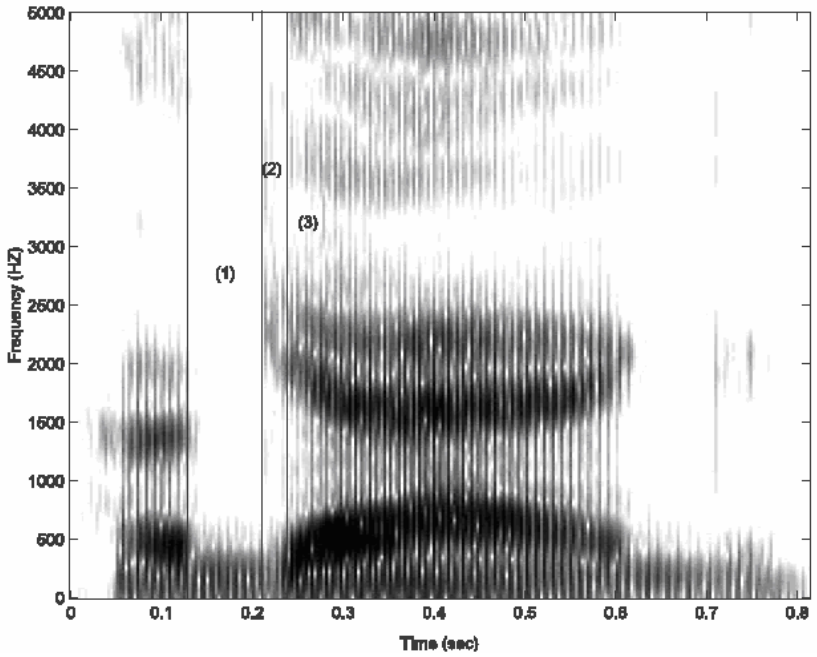

Figure 2-2 : A spectrogram of the utterance / g ae g/.

Three primary articulators, which are lips, tongue blade and tongue body, are used to utter different categorized stop consonants. For labial stop consonants,

the constriction is formed by the lips, for alveolar stop consonants, the closure is formed by the tongue blade and the alveolar ridge and the tongue body and the soft palate, or the posterior portion of the hard palate depending on the context of vowel, form the closure for velar stops. All of these three consonants will be explained in the following sections in details.

2.2.2. Un-aspirated Stop Consonants

The production of /b/, /d/ and /g/ belonging to this category are explained with acoustical events governing the nature of these sounds. It will be helpful to follow the articulation steps by inspecting the corresponding X-ray images for better understanding.

2.2.2.1. Un-aspirated Labial Stop Consonants

At the instance that the vocal tract is ready to utter a voiced stop consonant (/b/) proceeded by any vowel, the tongue body position is observed to be very close to that of the upcoming vowel utterance (See Figure 3-16). The major difference between the vowel and the plosive schemes seem to be the teeth and the lip openings i.e. in the labial and alveolar regions. Thus the formant trajectories after the release depend greatly on the following vowel, and the dominant part of the formant transition is caused by the motion of the lips and the jaw rather than the movement of the tongue body or the velar region. As a consequence, the tongue is almost stationary (Compare previous figure with Figure 3-17).

The movements of the articulators cause a change in the formant frequencies. Proceeding from labial release to a back vowel (/a/), F1 rises rapidly while there is a small upward movement in F2. F1 rises in a similar fashion when the following vowel is a front type (/i/), but F2 rises more slowly [10]. In order to check the validity of these observations, synthetic voices are constructed using Fant's tomographic area data [27]. In this construction the reflection coefficients and the areas are changed monotonically to provide continuity. The filter coefficients are obtained using state space model and the resonances of the filters are plotted in Figure 2-3 for two types of the vowels without any concern of the

accuracy but just to show the parallelism of the model with the nature. It is clear that the actual character of the transition is consistent with the model for labial releases if Figure 2-3 is compared with Figure 2-4 [10].

Continuing with the properties of the transition, it can be said that, the spectral shape of the burst is rather flat since the constriction is close to the opening of the tube [10]. Thus the spectrum of the burst is roughly the spectrum of the noise with smooth spectral shape (as modified by the radiation and excitation characteristics. With a rough approximation, the source can be modeled as a low pass signal with -12db per decade slop and the radiation can be modeled as a high pass filter with +6db per decade slope at high frequencies resulting a low pass characteristics approximately -6 db per decade). On the other hand when the stop is preceded by a vowel, the formant trajectories seem to be the symmetrical of the former case. Examples of spectrograms showing the formant trajectories of labial stops preceded and followed by front and back vowels are shown in Figure 2-4 (a) and (b) respectively [10].

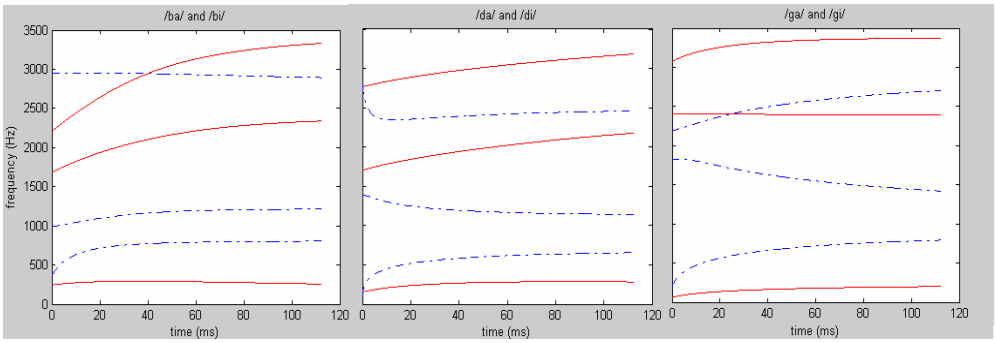

Figure 2-3 : Synthetic resonances of the vocal tract for three plosives. Solid lines correspond to the resonances of the transitions with front vowels and dashed lines are for back vowels

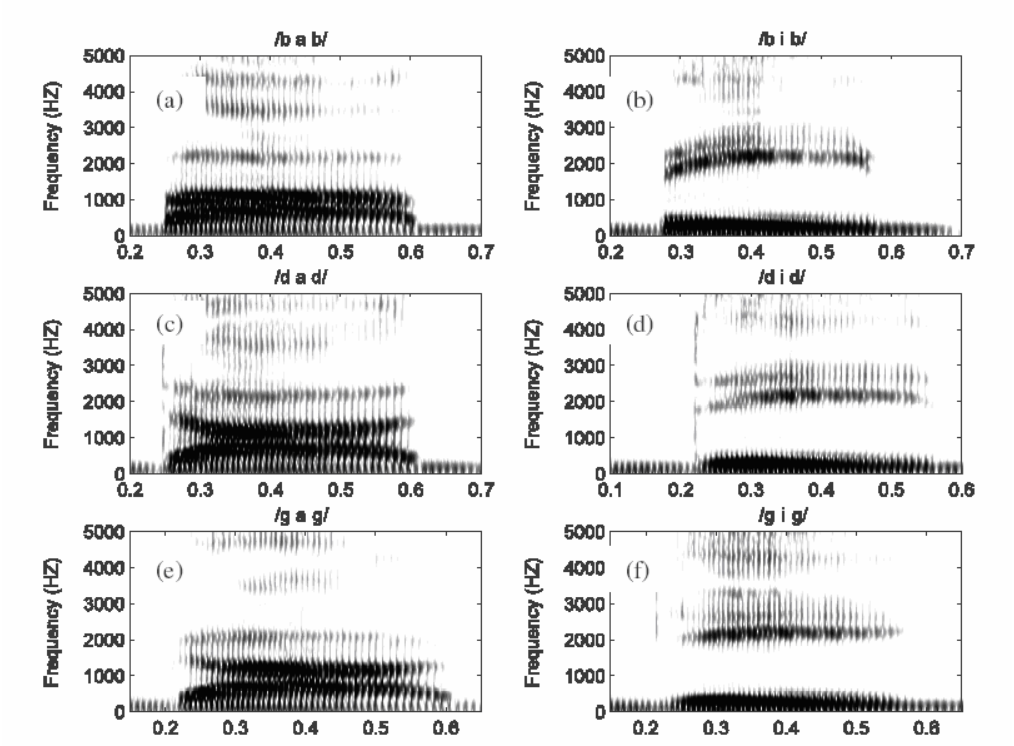

Figure 2-4 : Spectrograms of the utterances of (a) /b a b/, (b) /b i b/, (c) /d a d/, (d) /d i d/, (e) /g a g/, and (f) /g i g/.

2.2.2.2. Un-aspirated Alveolar Stop Consonants

/d/ belongs to this category of stops. For utterance of alveolar stops, the vocal tract is constricted at the alveolar ridge by the help of the tongue blade. Such a configuration needs a tongue body that is placed in a rather forward and upward position. Thus, the areas around pharynx and larynx are observed to be larger than natural configuration of the vocal tract (uttered vowel without any articulator manipulation is called natural vowel, see Appendix) expectedly. According to the acoustic tube modeling, second and third formant frequencies of the front vowels are higher than those of the back vowels. Therefore the second and the third formants are expected to decrease while progressing from alveolar stop to back vowel because alveolar constriction forces the vocal tract to look like a front vowel at the constriction stage. In the case of an alveolar stop followed by a front vowel, the tongue body at the constriction is almost at its vowel target for front vowel

generation (same reasoning with the back vowel case). Thus the tongue body generally moves slightly downward and forward into the position of the front vowel, resulting an increase in F2 according to the acoustic model (See Figure 3-18 and Figure 3-19). For both types of succeeding following vowels, F1 increases due to tongue body's downward movement (Compare Figure 2-3 and Figure 2-4) for the resonances of the actual transition and the synthetic one). In addition, the constriction between the alveolar ridge and the tongue blade accounts for a short front cavity with high resonance frequency, resulting in a burst spectrum with energy concentrating more in the high frequency region when the cavity is excited by the frication noise [10]. Examples of the spectrograms showing the formant trajectories of a front vowel and a back vowel preceded and followed by alveolar stops are shown in Figure 2-4 (c) and (d) respectively [10].

2.2.2.3. Un-aspirated Velar Stop Consonants

/g/ is the consonant of this category. For utterance of /g/, vocal tract is constricted from the soft palate by the back of the tongue if the upcoming vowel is a back type whereas the constriction occurs at the posterior portion of the hard palate for upcoming front vowels. The absolute difference between second and third formants decreases due to the constriction made by the tongue body and the soft palate compared with the difference between F2 and F3 of the uniform vocal tract (See Figure 2-4 (a) and Figure A-3 in the Appendix). The reason of this convergence can be explained using acoustic tube model such that, progressing from a back vowel /a/ (as in h<u>o</u>t) to the mid vowel /er/ (as in b<u>ir</u>d), velar alveolar and labial regions are narrowed resulting second and third formants to converge and first formant to decrease. (First three formants for /a/ are 730, 1090 and 2440, for /er/ they are 490, 1350 and 1690 [1]). The velar constriction can be thought as the further steps (extreme case) of this transition in which the convergence of these two formants is expected to be higher than the vowel case. For most vowels the configuration is similar, however, for a front vowel, the convergence of F2 and F3 is lower than in the case of a back vowel. This situation can be related to the vocal tract configuration for front vowels for which the vocal tract is already narrowed from the velar region. Thus, a major change i.e. a convergence is not

expected as in the case of back vowels. Instead, F2 and F3 increase from velar stop to a front vowel. F1 increases for all types of stops and all types of succeeding vowels because of the constrictions done by the vocal tract articulators. As a difference for velar stops, the movement of the formant is not as rapid as the ones in the alveolar and labial cases, since there is a greater length of constriction [10]. In addition to the greater length, the massive structure of the tongue body compared to the lips and the tongue blade also slows down the velar transition in contrast with other types [9]. Figure 2-4 (e) and (f) show two of spectrograms of the utterances /g a g/ and /g i g/ respectively. (See right part of Figure 2-3 and bottom part of Figure 2-4 to compare actual formants with the output of the concatenated tube model)

2.2.3. Aspirated Stop Consonants

Aspirated stop consonants /p/, /t/, and /k/ are the unvoiced versions of voiced ones /b/, /d/, and /g/ respectively. They do not differ from the voiced ones from the constriction side of view. The only difference is the voiceless excitation thus, the noise component of the transition. When unvoiced stops are progressed by vowels, aspiration noise is produced immediately after the constriction relaxation. Duration of the VOT is increased because of the aspiration process and also the closure of the glottis is delayed by the same reason. According to Atiwong [10] this is the main cause of aspiration noise and which is different from the un-aspirated case where the glottis remains in a more closed position with the vocal folds vibrating. As soon as the oral cavity closure opens, the pressure from the lungs drives the air to flow rapidly through the glottis opening, causing an airflow that acts as a noise source at the glottis. This noise source excites the vocal tract from the glottis to the mouth opening like a turbulence passing through the path. Thus, although the excitation is noise like, it is possible to notice the vocal tract resonances that are close to those of vowels at the aspiration noise period as the vocal tract moves from the stop to the configuration for the following vowel. The release bursts, like the ones in their un-aspirated counterparts, are still generated in the same fashion, and are superimposed with the aspiration noise. Furthermore,

the voice onset times (VOT) in aspirated stop consonants are typically longer for un-aspirated consonants.

2.3. Chapter Summary

In this chapter, the articulatory mechanism of stop consonants is described with discriminating events in a cause and effect relationship. The acoustic events involving in the utterance of these sounds are tried to be related to the acoustic concatenated tube model for consistency. The mechanisms are tried to be supported by the visual aids for better understanding. As a summary, the concatenated tube model is observed to be consistent with the theory for plosive to vowel utterance.

CHAPTER 3

CHARACTERIZATION OF PLOSIVE TO VOWEL TRANSITIONS

In Chapter 2 the events involved in the utterance of plosive to vowel transitions are explained with acoustical reasons. These reasons are tried to be related to the acoustic tube model. The purpose of this part is to analyze the plosive to vowel transitions in various aspects, more quantitatively.

In order to extract the properties, various examinations are done on plosive to vowel transitions. The examinations can be categorized mainly in four groups which are:

- Frequency domain examinations.
- Time domain examinations.
- X-ray examinations.
- Lips video examinations.

Frequency domain examinations are related with the spectral characteristics of the transitions such as the formant structure, burst spectra, aspiration character etc. Time domain examinations account for the properties that can be extracted from the time series of the speech such as pitch, power, timings etc. Finally the X-ray and lips video examinations, which are the main parts of this study, are the inspection of the movement of acoustical articulators. All parts give us some information about the character of the transitions thus the observations in different groups are tried to be related to each other. The perceptional properties are also mentioned with the experiments done.

In addition to the plosive to vowel transitions, fricative to vowel transitions are also analyzed with same aspects. The reason behind this decision is the existence of acoustic similarities between the fricatives and the plosives that are uttered by same movements of the articulators. In addition to similarities, differences are also figured out for a complete comparison.

Before proceeding, it will be helpful to review several studies which exist in the literature:

About characterizing acoustical properties of stop consonants, Halle et al. [23] and Fischer & Jorgensen [24] performed one of the first studies. Stevens [20] [21] [22] and Victor [9] were the followers which clarify many properties of these sounds. In recent years Jackson [4] [5] and Atiwong [10] also worked with stop consonants for clarification and examination purposes.

In parallel with acoustical properties Cooper et al. [14] investigated the perceptional properties of plosive sounds. Alwaan [34] proposed a model for perception of plosives in noisy environment.

On the vocal tract visualization studies, Story et al. [25] performed MRI investigations, Fant [31] and Stevens et al. [8] used tomography and X-ray respectively. As a special velar articulatory study, Perrier et al. investigated the vocal tract movements during velar constrictions [12].

With help of these studies Baer et al. [16], Narayan et al. [17], Perrier et al [11], Martin et al [15], Alain et al [13], Maeda [26] and Mermelstein [7] proposed relations which are used to model the vocal tract area functions making use of sagittal distances.

In this chapter Jackson, Atiwong and Victor will be referenced frequently for their statistical studies related to spectral and time based properties of stop consonants. Cooper et al.'s results will be used about plosive perception. MRI data of Story, X-ray data of Stevens and velar details of Perrier et al. will be made use for model derivation.

For conversion of X-ray measurements to area functions Mermelstein, Alain, Perrier and Martin will be used.

3.1. Plosive to Vowel Transitions

This part is about plosive to vowel transitions which is the main scope of this study. Several investigations, which are explained before, are done to extract the characteristics. The examinations are done both for voiced and unvoiced plosives for completeness.

3.1.1. Spectrogram Examinations

Spectrogram investigations are important for understanding the spectral aspect of the transition. In this work, several plosive to vowel transitions are recorded in both directions. The recordings are done in an office environment with Sony F-V120 dynamic microphone and the recording device is the Hp Laptop with Conexant sound card. The sampling rate is chosen as 16 kHz and the A/D quantization is done using 16 bits. Speech samples are recorded in the form of VCV in which V's represent the vowel of the same type whereas C is the consonant of the transition. The vowels are generally chosen as /a/ and /i/ as commonly used cases of front and back vowels for coverage. Plosives are used with all types including voiced and unvoiced versions. The TIMIT (Texas Instruments & Massachusetts Institute of Technology) database is also used to make comparisons with existing literature and diversity purposes.

3.1.1.1. Voiced Plosives

The spectrograms are constructed using the software "Praat" [18] with a Hanning window of length 20 ms. A proper overlap is also adjusted for continuous spectrogram structure. The samples are pre-emphasized with +6db per decade filter to strengthen the high frequency band. Six typical VCV recordings including labial, alveolar, and velar transitions are plotted in Figure 3-1.

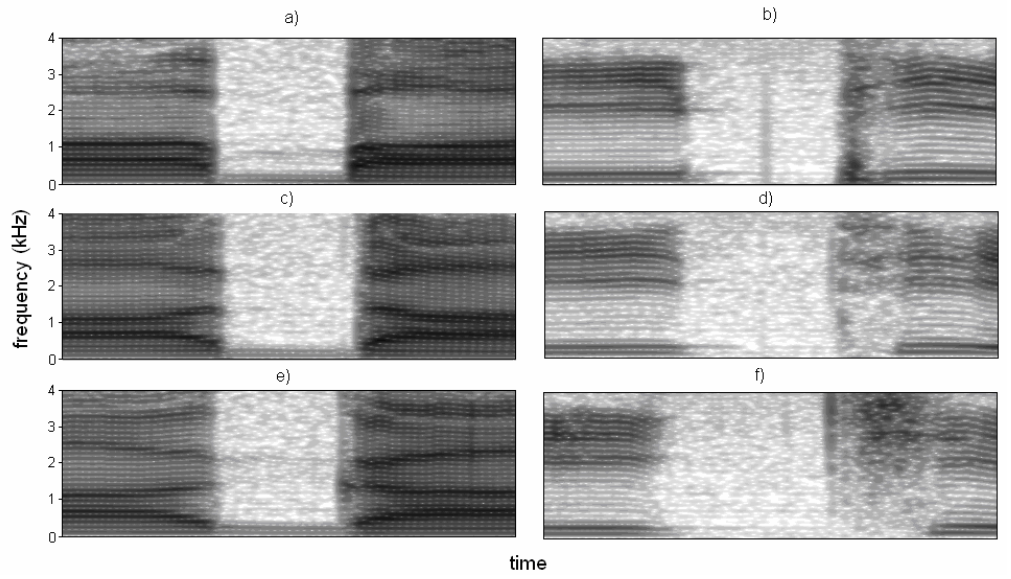

Figure 3-1 : Spectrograms of the utterances of (a) /a b a/, (b) /i b i /, (c) / a d a /, (d) / i b i/, (e) /a g a/, and (f) /i g i/.

First of all, the spectrograms seem to be consistent with Figure 2-4 of Atiwong's [10] and also obey the acoustic properties expressed in Chapter 2. Thus the similarity of the spectrums motivates us for a general characterization.

In the second chapter, the symmetric structure of both transitions i.e. the plosive to vowel and the vowel to plosive transitions are mentioned which are also illustrated in Figure 2-4 without emphasizing the reasons. According to the concatenated tube model, this observation can be related to symmetric movements of the vocal tract. More clearly, the idea of symmetric movement is the time reversal of the vocal tract movements during plosive to vowel transitions or vice versa. In addition to the symmetric structure, the smooth and continuous formant trajectories give us clues about the monotonic and the periodic utterance sequence in both directions. Moreover, around the burst parts, it was expected to have discontinuities at the formant trajectories, and having a flat like spectrum because of the impulsive structure in the time domain. But analyzing the burst part and the emphasized vowel to plosive transition regions, it is clear that neither of them contains such distortions. Symmetric, continuous, monotonic and simple formant trajectories even at boundary regions necessarily inform us about the importance of the formants thus the importance of smooth vocal tract movements

in these transitions. These observations also give us some idea about the dominancy of the vocal tract movements other than the plosive burst parts which give its name to these kinds of sounds.

The simplicity and smoothness of the formants forced us to analyze these trajectories in a more detailed fashion. For this purpose, the ensembles of /ba/, /da/ and /ga/ transitions are used. The formants of these ensembles are calculated with Praat [18] and all formant ensembles are plotted in Figure 3-2. The tracks of the first three formants are tried to be modeled with constant, linear, quadratic and exponential curves in the sense of root mean square error (RMSE, See Equation 3.1 for details, f_n and f_m denote natural and modeled formants respectively).

$$rmse\ f_i\ (Hz) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(f_{ni,\,j} - f_{mi,\,j}\right)^2} \quad i = 1,2,3 \tag{3.1}$$

For each model, the RMSEs (in Hz) of the formants are averaged over 30 ensembles and the results are compared with the ones of Jackson's [5] for labial stops (Other data is not available). Following table is obtained for comparison and verification.

Table 3-1 : The estimation results of the first three formants of the ensembles by constant, linear, quadratic and exponential models.

FORMANT	CONST.	LINEAR	QUAD.	EXP.	Time Constant
/ba/ F1	55.27	33.97	**18.09**	**8.13**	20 ms
/ba/ F2	30.97	16.02	**11.31**	**10.48**	
/ba/ F3	49.67	29.91	**15.67**	**16.7**	
/da/ F1	71.37	39.06	**13.39**	**8.98**	40 ms
/da/ F2	89.8	29.24	**18.24**	**19.66**	

/da/ F3	66.45	54.94	**45.27**	**53.3**	
/ga/ F1	81.35	27.42	**8.79**	**10.61**	80 ms
/ga/ F2	59.37	22	**14.03**	**15.23**	
/ga/ F3	84.69	27.71	**24.33**	**27.33**	

Table 3-1 : (continued).

Analysis of the table shows that, the exponential and quadratic models for formant tracks yield the least root mean square error for all kinds of plosives. Another interesting observation about the experiment is the best compatibility of labial plosives with exponential or quadratic tracks having small mean square errors. Indeed it is easy to figure out this observation just inspecting the ensembles. First formants of all types are exponential like having small RMSEs whereas for labial plosives the other formants also seem to have the same property. On the other hand it is examined that the RMSEs of the higher formants are larger than the first formants in general i.e. the RMSE increases as progressing from F1 to F3. In fact, this situation can be accepted as normal, as the average values of these formants as well as variances also increase in parallel with the RMSEs. To have a comparison, Jackson [5] found RMSE values for labial plosive to vowel transitions as 7.5 Hz, 8.7 Hz and 16.8 Hz for F1, F2 ad F3 respectively and we found them to be 8.13 Hz, 10.48 Hz and 16.7 Hz. It can be said that the measurements of ours and Jackson's are in an agreement for labial trajectories but unfortunately the results for other two plosives do not exist in his studies. However, the performance of exponential and quadratic tracks is obviously observed to be better than the others. Thus, it can be concluded that, the formant trajectories of plosive to vowel transitions are more likely to be exponential or quadratic rather than linear or constant ones. In this study the exponential structure is preferred for practical reasons.

In addition to the formant tracks, time constants of the transitions are also calculated from ensembles and are averaged. In some of the experiments, it is better to use a single time constant denoted by $\hat{\tau}_g$ for all formants to simplify the

computation. So it would be beneficial to define a general time constant. General time constant is obtained from the maximum of three time constants:

$$\hat{\tau}_g = \max\left(\tau_{f1}, \tau_{f2}, \tau_{f3}\right) \quad \textbf{(3.2)}$$

Where τ_{fi} s are the time constants of the formant trajectories. $\hat{\tau}_g$ is used as the general time constant if its variance is sufficiently small. In case of large variation in $\hat{\tau}_g$, the second largest time constant is selected as $\hat{\tau}_g$. Fortunately, in our calculations the maximum time constants for all transitions had small variances (F3 for /ba/ and /ga/ and F2 for /da/).

As a result, the time constant of the velar transition is selected as the largest of the three. Fastest formant changes during the transitions are for labial releases whereas velar stops are realized to be the slowest of the three. The calculations turn out to be in an agreement with the theoretically expected ones i.e. the mass of the articulator is directly proportional with the time constant [9]. Knowing that, the, mass of the tongue body is greater than the tongue blade which is greater than the lips, the time constants found become consistent with the literature.

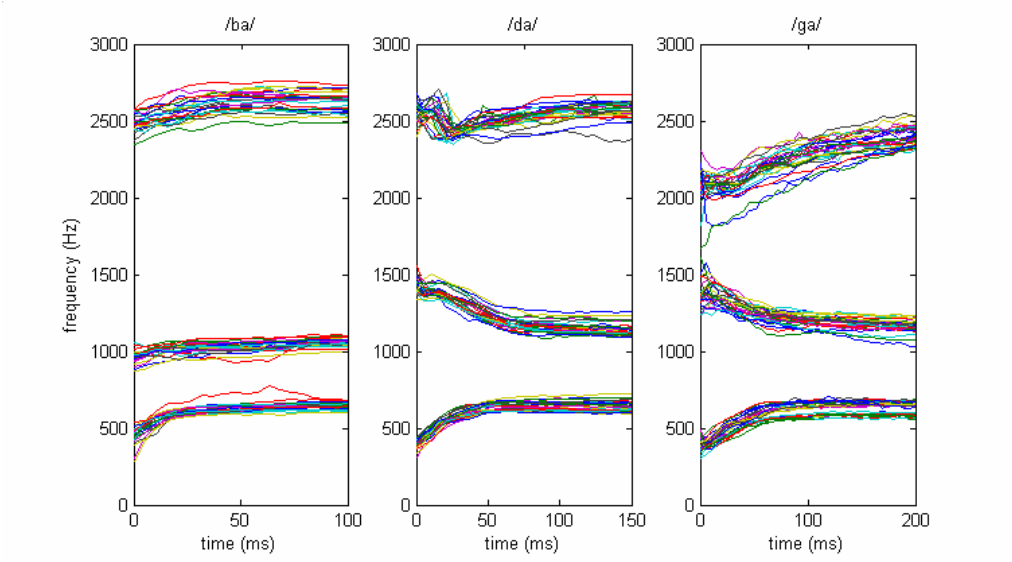

Figure 3-2 : First three formant frequencies of the ensembles with back vowels.

For quantitative analysis and comparison purposes Table 3-2 is constructed with ensemble data (Jackson's data is used for /i/ transitions).

Table 3-2 : The average of the formants of the burst and the steady state vowel

Transition	F1 (Hz)		F2 (Hz)		F3 (Hz)	
	Burst	Vowel	Burst	Vowel	Burst	Vowel
/ba/	442	642	960	1058	2478	2614
/da/	400	643	1423	1146	2538	2574
/ga/	392	633	1362	1165	2074	2400
/bi/	320	340	2100	2355	2416	3220
/di/	320	340	2089	2355	2765	3220
/gi/	150	340	2258	2355	3111	3220

Table 3-2 reveals that the plosive spectrum is highly affected from the vowel's spectral characteristic. The dominancy is even clearer for the second formant. Average second formants of /b/, /d/ and /g/ in the context of /a/ are found to be 960, 1423 and 1362 respectively whereas the same values in the context of /i/ are calculated to be 2100, 2089 and 2258 by Jackson [5]. These results give clear clues about the transition which starts with a spectral configuration very similar to those of vowel's. On the other hand, to compare the observations with the output of the primitive model, it will be better to analyze Figure 2-3 of exponentially changing synthetic formants and Figure 3-2 of actual formants with Table 3-2. Roughly, the basic model obeys the dominancy of the vowel and also the synthesized formant tracks for back vowels seem to be consistent with actual ones.

To sum up, formant tracks of voiced plosive to vowel transitions are observed to be smooth exponential tracks symmetrical in both directions. In addition to the symmetricity, it is noted that the behavior of the tracks are governed by succeeding vowels. These two properties are preserved in the primitive model

with some exceptions which can be related to the simplicity of the model and measurement errors.

3.1.1.2. Unvoiced Plosives

Like their voiced counterparts, the unvoiced plosives are also examined for comparison and completeness. Six unvoiced plosive to vowel transitions with the same configuration of the voiced case are shown in Figure 3-3.

Even with a rough examination, the symmetric structure of the formants and the similarity of the tracks with their voiced versions are easily observed. On the other hand the degradation of the smoothness at formant trajectories especially for the first formant, narrower formant bandwidths, high frequency components at aspiration and impulsive spectrum at the burst parts are the main differences which can be easily observed.

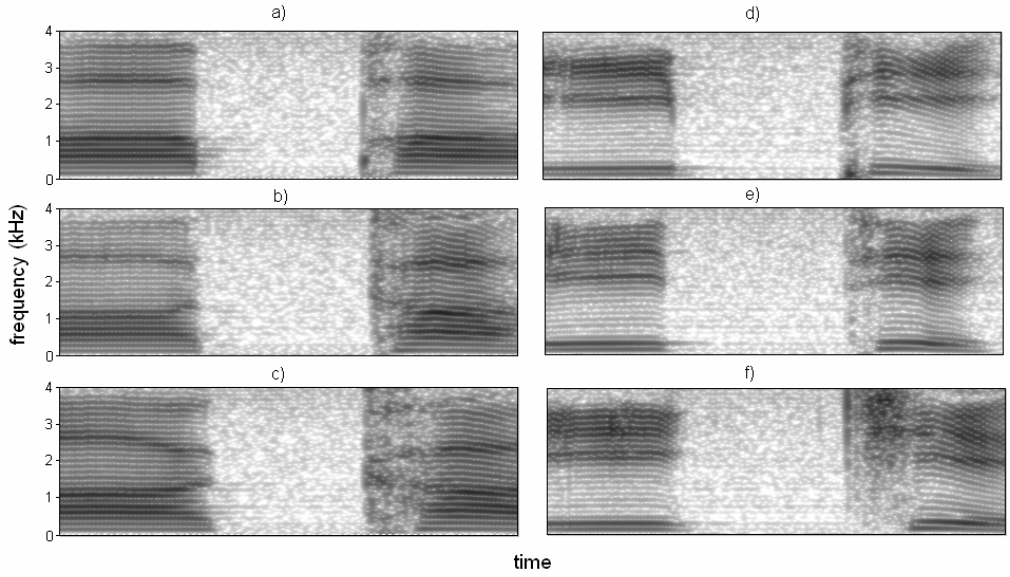

Figure 3-3 : Spectrograms of the utterances of (a) /a p a/, (b) /i p i /, (c) / a t a /, (d) / i t i/, (e) /a k a/, and (f) /i k i/.

The differences can be discussed in a comparative manner. The bandwidth and smoothness differences of the formants for voiceless plosive to vowel transitions can be related to the existence of an unvoiced excitation source from the beginning of the release till the generation of the upcoming vowel. The distortion effect is mostly observable in the low frequencies especially at first formant because of absence of low frequency excitation [9] In addition to the noise like excitation, aperiodicity is one of the main factors which decrease the power of the first formant together with low frequency components in the aspiration part compared to voiced case. On the other hand having a low pass excitation together with a periodic structure, these distortions are not expected in voiced plosives. The impulsive structure at burst parts of the unvoiced plosives which is absent for voiced ones can be related to the same reason.

However, despite of all these differences, within a rough manner, unvoiced plosives have many common properties with their voiced versions. First of all similar structure of the formant trajectories for the voiced and unvoiced excitations give some clues about the similarity of vocal tract movements during the transition. Moreover, it will be shown following parts that the fricatives of the same type have approximately the same vocal tract movements during transitions in. Secondly, the effect of the vowel at the aspiration part will be discussed and finally the character of the burst parts will be compared with voiced counterparts.

In order to examine the influence of the vowel to the aspiration part and also learn about the vocal tract state during the aspiration, an experiment is performed on /pa/ transition. In this experiment the lpc model of order 20 of the vowel /a/ is obtained from the steady state region of /pa/ transition. The AR filter is obtained by the extracted lpc is excited with a Gaussian noise of length 50 ms. The perception of the synthesized waveform by 5 subjects are found to be % 100 satisfactory. Also the formant structure of the transition is almost like the natural sound except some differences at the regions that correspond to the vocal tract movements since the synthetic one is filtered with a stationary model. For alveolar and velar cases the synthetic stationary filter output may not have good performance because of high vocal tract resonance differences. See Figure 3-3 a, b and c). This experiment shows the importance of the following vowel for unvoiced plosive to vowel transitions. As a result it can be said that the aspiration part is nothing but

the excitation of quasi-stationary vocal tract with the white Gaussian noise whereas the excitation is the pulse train for voiced transitions.

To demonstrate the effects more clearly, it will be helpful to analyze the experiment of Jackson's [5]. In Jackson's experiment, several labial, alveolar and velar transitions with both back and front vowels (/a/ and /i/) are recorded. Then the burst parts of these transitions are manually extracted with a Hanning window of length 6 ms. Extracted samples are padded with 24 ms of zeros resulting 30 ms signals. The power spectrum densities are obtained for each ensemble and are plotted in same graph for each transition. For statistical purposes, the ensemble averages together with estimated formant frequencies [5] (for voiced case) of the spectral densities are also calculated and Figure 3-4 is obtained as a result.

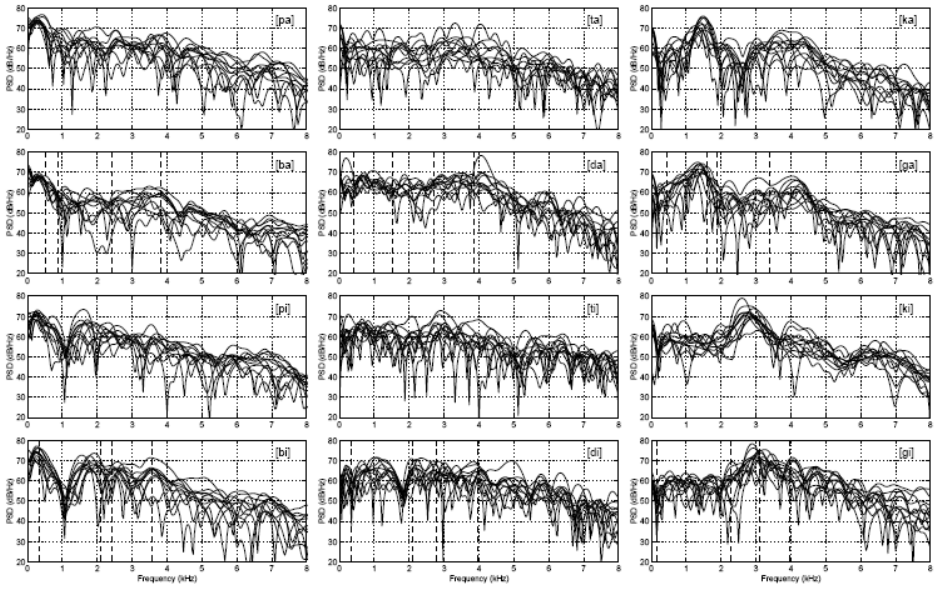

Figure 3-4 : Power spectral densities of the plosives in /a/ context (upper half) and /i/ context (lower half). The upper row within each half accounts for unvoiced plosives whereas the lower row accounts for voiced plosives. The thick lines represent the ensemble averages. The dashed lines are the predicted formants.

Examining the burst PSDs, it is easy to observe that not only the aspirated and un-aspirated transition (for unvoiced and voiced plosives respectively) but spectrums of the bursts are also under the dominancy of the upcoming vowels.

The dominancy shows itself especially at frequencies for which the vowel has its second and third formants. Thus, transitions with /a/ have their energy concentrated between 1 – 2 kHz whereas 2.5 – 3.5 kHz band is the most powerful interval for plosives in the context of vowel /i/. Analyzing separately for each kind of plosives, labial bursts are more likely to have low pass characteristics remembering a guess of -6db per decade in Chapter 2.2.1 [10] which is actually very similar to the real case except for some second and third formant effects of the following vowel anyhow. Then continuing with alveolar bursts, the accumulation of the energy in the band 3 – 5 kHz makes them more likely to be high pass (See part 2.2.2.2.). This situation can be related to the increase of the second and third formants for back vowels which results a high frequency part in the burst, whereas for the front vowel case the frequencies of second, third and fourth formants are already high so the corresponding formants of the burst part are powerful either. And finally for velar stops the situation is a little bit different from the others. Remembering the transitions, for both types of vowels, the second and the third formants were trying to converge to eacother. Thus, the spectrums of the bursts also have their energy mostly concentrated just between the second and the third formants of /a/ and just about the second and third formants of /i/ which can be accepted as a band-pass character for velars. These observations are also in an agreement with the literature [23].

A wider study on this subject is Victor's [9]. In his study many speech samples of velar and alveolar stops proceeding with many vowels are recorded. The center frequencies of the burst parts that correspond to a specific vowel are calculated and averaged over the samples. Following figure is a reedited version of his studies.

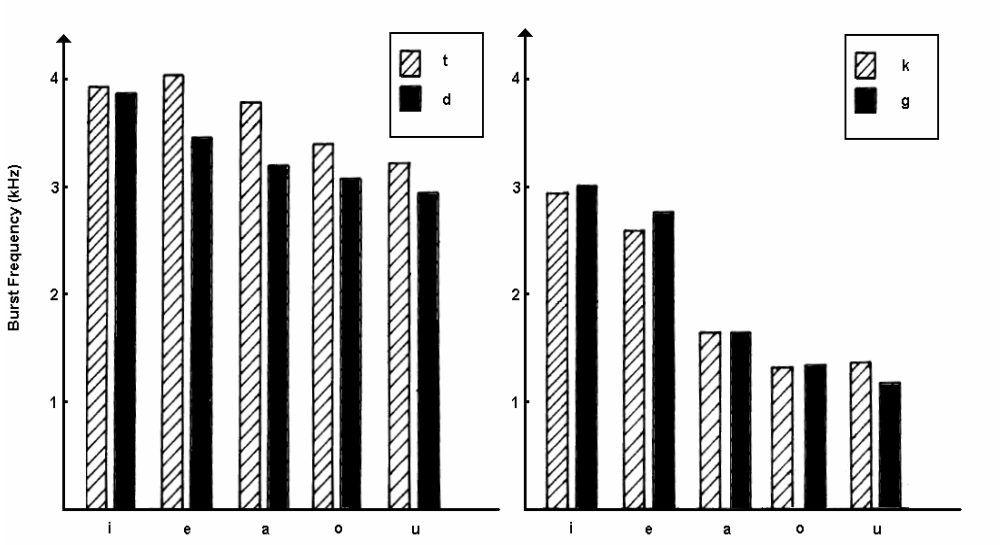

Figure 3-5 : The center frequencies of the bursts followed by various vowels.

Alveolar stops are observed to be in 3 – 4 kHz band depending on the vowel context being consistent with Jackson's [5] studies. Velars are in a mid band compared to the alveolars depending greatly on the following vowel (In Chapter 2 this property was related to the major vocal tract changes during the transition). The most important observation from Victor's and Jackson's studies is the resemblance of the burst spectrums of voiced and unvoiced plosives of the same type. To explain more clearly, the spectra of /b/ followed by /a/ is more similar to /p/ followed by /a/ than to /b/ followed by /i/. Thus, the upcoming vowel i.e. the vocal tract configuration is the main factor that governs the character of plosive to vowel transitions even at the starting point of these transitions.

As a last step, to intensify the vowel plosive relation from perception side of view following study of Cooper et al. [14] can be examined. In this study, 30 listeners are subjected to unvoiced bursts following several vowels and they are asked to recognize the bursts. The vowels are /i/, /e/, /ɛ/, /a/, /□ /, /o/ and /u/. The center frequency of the burst is changed from 360 Hz to 4320 Hz with 360 Hz steps for each vowel then the following figure is obtained with the information from the listeners saying whether the transition involved is /p/, /t/ or /k/.

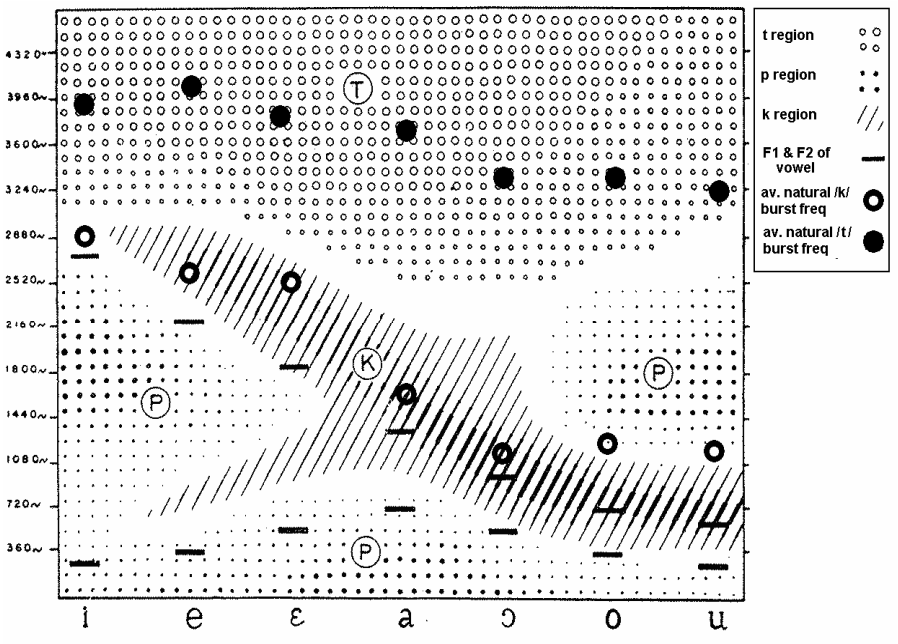

Figure 3-6 : Recognition rates of the synthetic plosive to vowel transitions.

Two thick lines in the Figure 3-6 stand for two characteristic formants frequencies of the vowels. Victor's data [9] is also superimposed on the graph for comparison purposes. The x axis is the vowel axis and the y axis is the center frequency of the burst. As expected, the high frequency bursts are recognized as /t/ whereas low frequency bursts are recognized as /p/. For /k/ (The observations are also consistent with Jackson's experiment) the bursts with center of frequency little higher than the second formant of the upcoming vowel are recognized as /k/. This study implies that the center frequency of /k/ followed by a vowel should have a center of frequency larger than the second formant frequency of the following vowel. This is because of the convergence of the second and the third formants to each other which is related to the vocal tract movements for velar releases. So it is expected to have a little higher center frequency of the burst than the second formant frequency of the following vowel for all vowels which is shown in the right part of Figure 3-5.

To sum up this section, it can be said that, the voiced and unvoiced transitions are very similar from the formant side of view. /b/ and /p/ are more likely to be low-pass, whereas /t/ and /d/ have their energy concentrated at higher

frequencies depending on the vowel context. /k/ and /g/ depends greatly on the following vowel having their center of burst between second and the third formant of the vowel. The importance of the vowel context reveals the importance of vocal tract movements other than any other factor involve in the utterance of the transitions.

3.1.2. Time Waveform Examinations

3.1.2.1. Voiced Plosives

Waveform investigations are essential to figure out some other properties of transitions such as intensity, pitch period etc. Therefore the same samples with back vowels, which are analyzed in the previous section, are also used in time waveform analysis. Following plots shows typical transitions that will be used in this part.

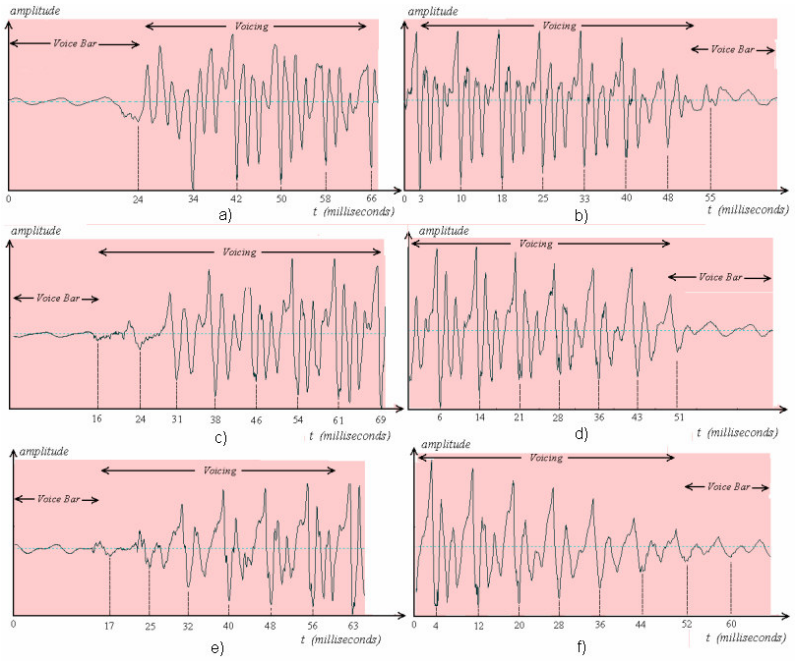

Figure 3-7 : Time waveforms of the utterances of (a) /ba/, (b) /ab/, (c) /da/, (d) /ad/, (e) /ga/, and (f) /ag/.

From Figure 3-7, it is clear that, the periodicity is preserved in both directions having no complex burst like structures in any direction of the transitions. As a small exception a 10-15 ms interval of noise is observed to be superimposed at the starting point of /ga/ and /da/ utterances which denotes VOT region but this region does not violate the periodicity. Again, the reason of this observation can be related to the smooth, continuous and monotonic movements of the vocal tract. As explained before, when the vocal tract is started to be constricted, high frequency components of the voiced speech start to decrease, thus the waveform starts to look like the voice bar until the total closure occurs. During total closure, the excitation radiates from the soft skin and the speech waveform takes the place of voice bar instead of resembling to it. In the opposite direction as the vocal tract starts to be relaxed, the excitation starts to get filtered by the vocal tract and the high frequency components of speech are increased because of the vocal tract resonances. Normally, a burst is expected at the starting of the relaxation interval but it is not essential for the characteristics of the transition. As times goes by, the relaxation effect decreases continuously forming the vocal tract ready for the upcoming vowel, thus the waveform starts to look like the vowel waveform in a monotonic fashion, completing the transition. Thus it can be concluded that the transition is in a very periodic manner without any distorting components that violates the periodicity even at the burst parts. To examine the periodicity of the transition part, the ensembles that are used before, are analyzed with Praat [18]. Following figure is obtained for the pitch period for three plosives and two vowels including the VOT regions.

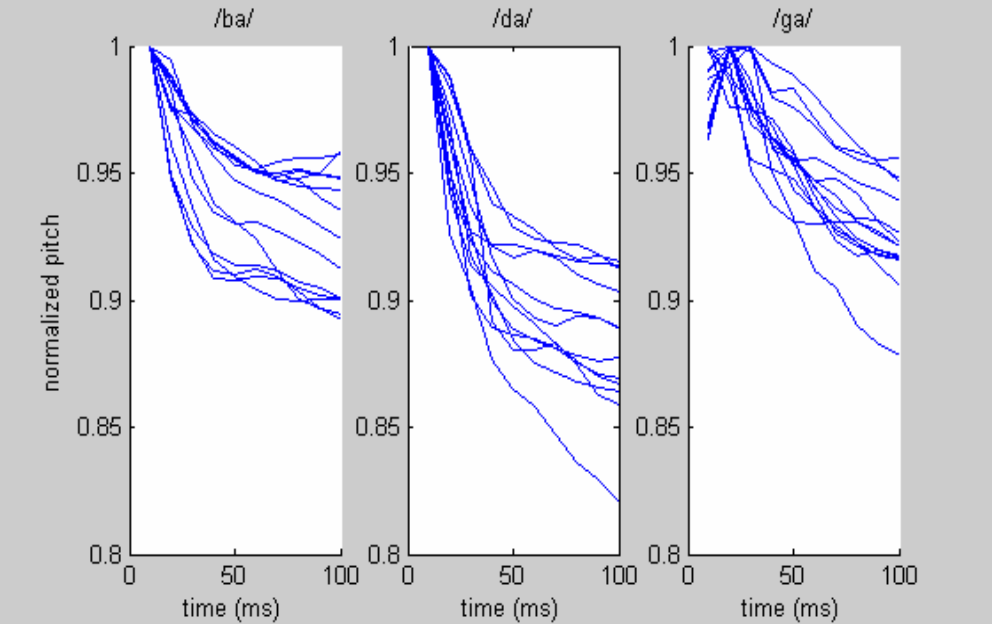

Figure 3-8 : The ensembles of the pitch values for /ba/, /da/ and /ga/ transitions.

From Figure 3-8 it is observed that for all kinds of plosive to vowel transitions, the pitch period seems to be decreasing with an approximate mean value of ten percentage and the decrease can be modeled by an exponential curve. As it is told before the pitch period of the burst parts are also tried to be calculated with the software. The percentage of successful pitch tracking for burst parts was more than ninety percent which is a reasonable percentage to accept the burst parts as a part of the periodic transition.

Table 3-3 : Average VOT values of plosives in the context of various vowels and the standard deviations

Plosive	B	P	D	T	G	K
VOT (ms)	13	58.5	19	69.5	30	74.5
STD DEV. (ms)	3	14	3	7	5	11

In fact, the reason of the high percentage of the successful pitch tracking can be related to the short duration of the VOTs for voiced plosives. For an idea and a comparison with the unvoiced ones, Victor [9] calculated the average length of VOTs for voiced and unvoiced plosives for all types. He also calculated the standard deviations for each transition. The results are summarized in Table 3-3. The VOTs for unvoiced plosives are about 3-3.5 times larger than voiced plosives. The VOTs for voiced plosives are very short especially for labial ones (even they could not be observed most of the time for labials and alveolars) and expectedly the velar case is the longest for the reasons explained before.

The periodicity in time domain and the continuity of formants with respect to time (spectrum analysis) increased the curiosity about the perception mechanism of the plosive sounds. Thus, in order to explore the importance of the VOT region and the formant differences in the perception of these transitions an experiment is done. For the experiment, different aged 25 Turkish subjects are asked to distinguish the labial, alveolar and velar releases followed by vowel /a/ in six experiment sets. In each experiment set, different transitions for three different plosive types (randomly) are asked to be recognized. But, for each set one period of these sounds are clipped. The first set contains the original sounds and the last set contains five period clipped sounds. Each sound consists of /xa/ transitions (x denotes a plosive) and subjected to subjects in a silent office environment using Sony MDR-P180 headphones. The average of the recognition rates of the subjects are illustrated in the following figure.

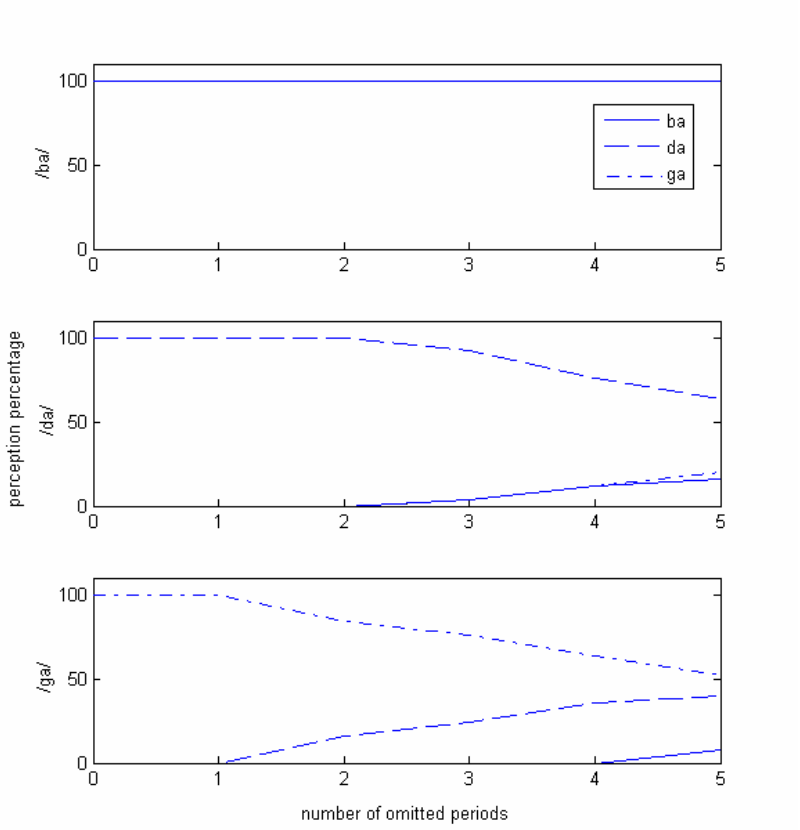

Figure 3-9 : Perception rates of clipped transitions.

In this figure each row corresponds to /ba/, /da/ and /ga/ truncations respectively. In each part of the figure x axis corresponds to truncation number i.e. the number of omitted periods and y axis corresponds to the recognition rates. Thus, summation of all rates in a vertical line in each part of the figure equals to 100 percent as expected. Each plosive recognition rate is plotted as indicated in the legend to resolve ambiguity.

The recognition rates of /ba/ and /da/ transitions do not seem to be effected by clipping the data even with a high amount of clipping (Typically 20-30 ms if a typical pitch of 120 Hz is considered for male, further truncation do not decrease the recognition rates but the transition turns out to be a single vowel). However, velar recognition rates decrease more rapidly. Besides, the wrong recognitions, velar transitions are noted to be confused generally with alveolars. The reason of

the confusion of velar releases with alveolars and also successful recognition of labials and alveolars despite of high amount of truncation can be explained with another experiment done by Cooper et al. [14].

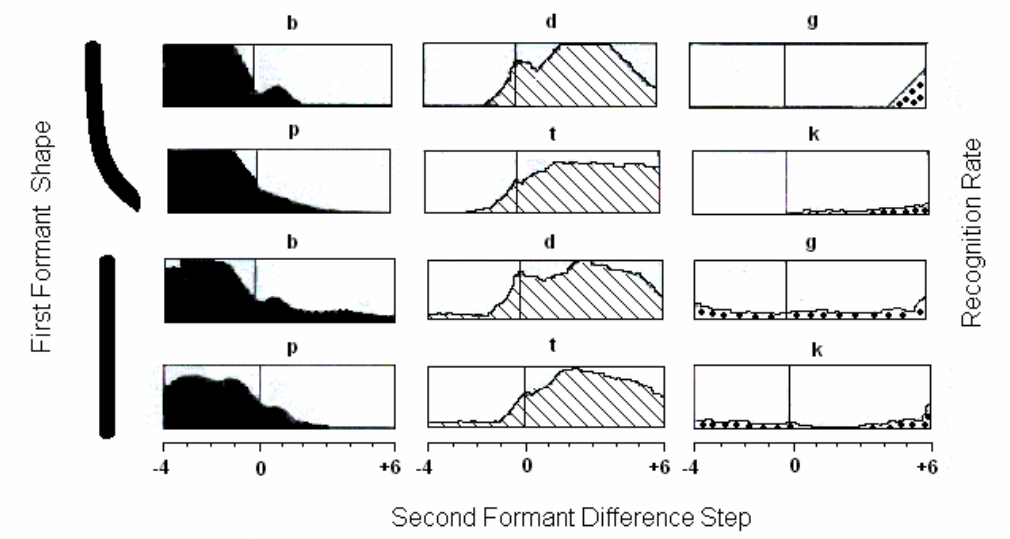

Figure 3-10 : Recognition rates of synthetic plosives in the vowel context of /a/ with different formant configurations.

In this experiment 30 people are asked to recognize synthetic voiced and unvoiced plosive to vowel transitions in two main sets. In the first set, progressing from synthetic plosive to vowel, first formant of the plosive is set to be properly smaller that the first formant of the vowel (as in the case for all natural plosive to vowel transitions) simultaneously the second formant of the plosive is changed gradually in 10 steps with respect to the vowel. The result of the experiment is shown in Figure 3-10. X-axis of the graphs represent the relative frequency of the second formant. More clearly, +6 corresponds starting formant frequency at the beginning, 6 level higher with respect to the vowel. Subjects are asked to recognize these synthetic sounds whether they are labial, alveolar and velar transitions. In the second set, the configuration is the same except for the first formant which is kept constant during transition. Recognition rates are recorded and finally Figure 3-10 is completed with a probability density function type

approach (In Figure 3-9 actually all columns are plotted in a single part whereas Figure 3-10 shows them separately as three parts).

First of all, for each set and for each voicing type the recognition rates seem to be very similar. Therefore it can be deduced that the perception mechanism of voiced and unvoiced plosives are similar under the dominancy of the second formant frequency, Cooper et. al [14] (The spectral similarity of /ba/ and /pa/ or any other couples also gives clues). In more detail, if first row of the Figure 3-10 is analyzed (actually the row which resembles the natural counterparts mostly), it is observed that with an increasing first and second formant frequencies, the synthetic sounds are percept as /b/. On the other hand as the difference of the second formants of the plosive and the vowel is decreased, the sounds start to be partially percept as /da/ and /ba/. Further increasing of second formant beyond the second formant of vowel yields a perception of /da/ alone. If difference is kept increasing recognition rate of /da/ dramatically decreases and the sounds start to be percept as /ga/. The results are in an agreement with the natural formants of these stops (See Figure 2-4) and they are also consistent with our experiment done by natural sounds. Consistency will be interpreted more clearly if Figure 3-11 is analyzed.

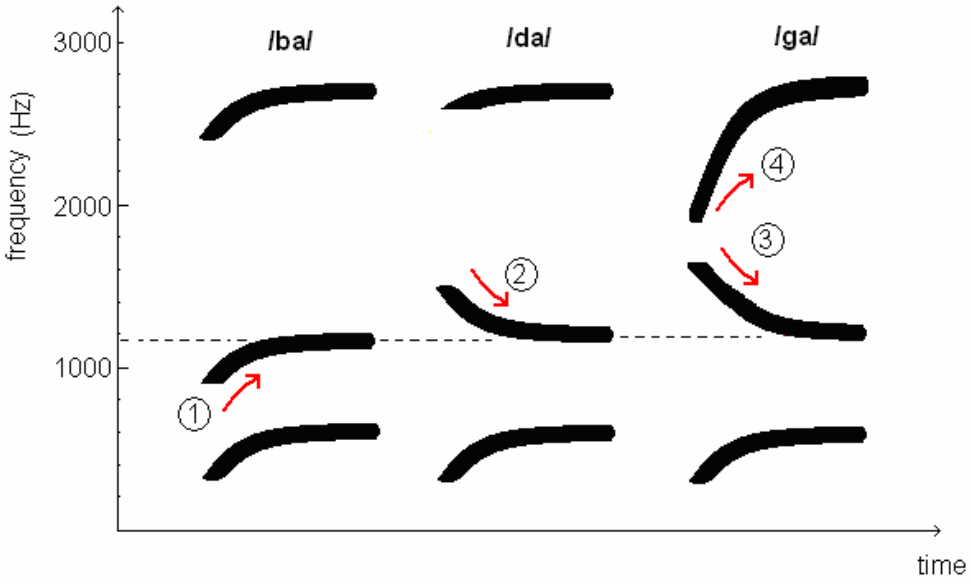

Figure 3-11 : Formant tracks of utterances /ba/, /da/ and /ga/.

This figure shows first three formant trajectories of three plosive utterances in the context of vowel /a/. Formant values of the start and end points of the trajectories are obtained from Jackson's ensembles [5] discarding the time information for this case. Returning back to wave truncating experiment, as the starting of /ba/ is truncated in time domain, all formants are also truncated in the direction of the arrow numbered 1 in frequency domain. The recognition rate is not expected to decrease as there is no other alternative utterance which has increasing second formant during the transition. Note that, Figure 3-10's first row shows that, almost all plosives, which have F2 lower than /a/'s second formant, are recognized as /b/. Further truncation of /ba/ without equalizing start and end points of the second formant did not yield /ba/ and /da/ confusions and it is also possible to observe similar results from Cooper et al.'s experiment [14]. For /da/ transition, data loss caused a spectral loss in the direction of the second arrow. The loss did not introduce confusion to recognition at first but as the loss is increased, again /ba/, /da/ and /ga/ confusions are encountered as in the case of Cooper's. Finally /ga/ transitions are recognized correctly at first truncations but later, confusions are observed as the spectrums are introduced loss in the direction of the arrows 3 and 4. This is an expected result as two of the spectrums have common properties in the first two formants. In Figure 3-10's first row, /da/ recognition seems to be dramatically decreasing as the second formant frequency of the synthetic plosive is increased. On the other hand, as /da/ recognition is decreased, /ga/ recognition seems to increase at the same rate. However analyzing Jackson's [5] ensembles starting formants frequencies of the /da/ and /ga/ transitions seem to differ about 100 Hz which is not a marginal difference to affect the recognition of /ga/. Although Cooper et al. [14] assumes that the second format is the dominant factor which governs the recognition of the plosives, for velar releases in the context of /a/ the importance of the third formant exists which he did not account for. Examining Figure 3-10's first row again, it is clear that increase of second formant improves the recognition of /d/ but further increase gives a rise to /g/ whereas /d/ starts to decrease. Because of the context of the study, it can not be progressed more but it can be guessed that further increase in the second formant totally decreases the /d/ rate and /g/ is totally recognized. Therefore it can be deduced that, converging third formant towards the second

formant is effective for the recognition of /ga/ transitions because of the cumulative boost at frequencies around these two (See also Figure 3-6).

To sum up, it is observed that periodicity and the symmetric structure is almost preserved at the time waveforms. No burst like structure is observed that inhibits the periodicity. It can be concluded that, during the transition the periodic excitation continuously excites the system while the vocal tract smoothly transforms its shape for the following sound. If the following sound is a vowel, the vocal tract relaxes the constriction to form the shape of the vowel. On the other hand if the upcoming sound is a plosive, the vocal tract starts to constrict itself from a proper articulation place transforming its shape forming the shape for plosive utterance. On the perception side of the view of these transitions, it is observed that the VOT region is not critical for labial and alveolar stops whereas for labials its impact is a little bit more but still the recognition of the velars are not full percentage dependent on the VOT region, instead it is observed that the formant differences between the starting and the finish points are also important.

3.1.2.2. Unvoiced Plosives

Time waveforms of the aspirated version of /b/, /d/, /g/, which are /p/, /t/ and /k/, are plotted in Figure 3-12. Comparing Figure 3-7 and Figure 3-12, completely different characters are observed. The reasons of the difference are simply the aspiration component, which is absent in voiced ones, and the length of the burst parts, which is generally unobservable in voiced transitions. Thus, up to voicing part, no periodicity is observed. Because of the absence of periodicity, the pitch concept can not be argued even. Instead, aspiration plus burst parts can be argued (VOT region). From Table 3-3, velar VOT is seen to be approximately the same with alveolar one which are greater than the labial VOT. The same configuration can be observed from the Figure 3-12 in which the velar VOT is found to be 92 ms. and the alveolar VOT is the nearly .Unlikely, the labial VOT is observed to be 64 ms.

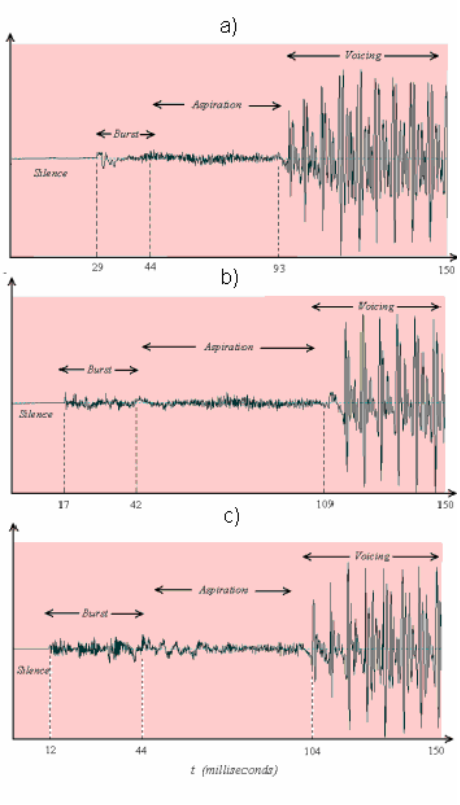

Figure 3-12 : Time waveforms of the utterances of (a) /pa/, (b) /ta/ and (c) /ka/.

One more difference between the voiced and unvoiced cases is the silence part that takes place at the beginning of the transition. As told before, the reason for voice bar which exists in voiced transitions is the existence of the voiced excitation that excited the soft vocal tract walls. The unvoiced transitions are excited by unvoiced excitations including the burst part. Therefore after the fist vowel utterance the voiced excitation is stopped for the burst part and the voiced bar is not formed.

To sum up, the periodic structure in voiced transitions are not observed for unvoiced plosive to vowel transitions, instead long aspiration and observable burst parts are noticed. Thus the VOTs are calculated to be much longer compared to the voiced plosives and pitch values can not be defined up to voicing part as expected.

3.1.3. X-ray Examinations

Previous sections account for the analysis of the outputs and the properties that can be derived from the outputs. A highly complicated production of the transition can not be precisely examined only by examining the output of the system. Thus, to explore the low level acoustic events involved in the production of the plosives and the differentiating properties of these sounds, they are needed to be analyzed by vocal tract visualizations in addition to the output properties.

In the literature there are many methods to visualize the vocal tract, like Magnetic Resonance imaging known as MRI, X ray imaging, Ultrasound imaging, etc. Depending on the technology that the method uses, the capabilities differ.

In this work, we have used a video from Ohman and Stevens [8]. This video contains X-ray images of the vocal tract of a subject. The subject is a 38 year old Canadian native male and the video is recorded in 1962 at the cineradiographic facility of the Wenner-Gren Research Laboratory at Northull's Hospital in Stockholm, Sweeden. The images are simple bitmap images sampled at 30 Hz hence the interval between two frames is approximately 33.3 milliseconds. In addition to snapshots, the video stream also contains an audio stream which is sampled at 10 kHz and quantized with 8 bits.

We focused on labial, alveolar and velar transitions in context of front and back vowels. Voiced and unvoiced categorization is not handled because of the almost exact vocal tract movements during production of both types (Story et al. [25], Shrikant et al. [17]).

3.1.3.1. Velar Snapshots

In Figure 3-14, a velar plosive to back vowel transition is figured out by the help of continuous video snapshots. The region of the vocal tract is covered with red lines to track differences easily in each snapshot. Here, the dashed lines should be ignored, later the sagittal distances will be calculated with their guide. In the first snapshot, the vocal tract is about to relax the velar constriction which is done by the soft palate and raised tongue body. Inspecting the exact constriction place, the cavity opening at the mouth, the tongue and the vocal tract indeed gives

clues about the upcoming back vowel. Inspecting one by one, the constriction place is observed to be at the soft palate whereas it would be at posterior portion of the hard palate if the upcoming vowel is a front vowel as illustrated in Figure 3-15. This is an expected result because as the vowel type changes from back to front (mid vowel between) the tongue is kept in a more forward position. Thus body of the tongue also constricts the velar region from a more forward region (See Figure 3-13 [12]).

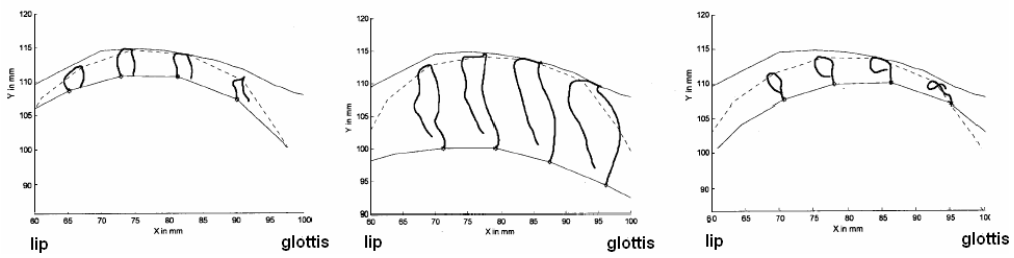

Figure 3-13 : Virtual velar trajectories of /iki/, /aka/ and /uku/.

In Figure 3-13, the palate (upper curve) and the tongue body (lower curve) are represented by two thin curves that are referenced to a fixed point in front of the mouth. The thick lines represent the trajectories of four points on the tongue body which are obtained in a simulative environment that is consistent with muscular model of the tongue and also the literature data describing the tongue movement in which the vocal tract transits from one of vowels /i/, /a/ and /u/ to a the velar plosive /k/ and again to the same vowel. The constriction point slides from the hard palate to the velum as mentioned before.

Continuing with Figure 3-14, in addition to the constriction region, the large sagittal distances observed at the front palate, alveolar ridge and the lips despite of the constriction that forces to straiten all these sagittal distances are the signs of the incoming back vowel /a/. Unlike these large widths, the narrowness of the pharynx and the larynx regions because of the back withdrawn tongue root are other clues for the vowel.

As the slides pass, the root of the tongue is more withdrawn to the pharynx, narrowing pharynx and larynx regions more. The tongue body is lowered for back

vowel utterance. The tip and the blade of the tongue are also lowered for enlarging alveolar ridge area. At last, the lowering and withdrawing of tongue is completed with widened alveolar ridge by the lips and the tongue tip which make the vocal tract ready for /a/. A detailed comparison of two vocal tract snapshots, in which a front vowel and a back vowel velar transition exist, can be obtained by comparing Figure 3-14 and Figure 3-15. In Figure 3-15 the constriction region is obviously different from ka/ga case as mentioned before. Moreover, the vocal tract configurations are completely different. The details of the back vowel transition are mentioned before. For the front vowel /i/, the tip and the blade of the tongue are elevated in an opposite manner compared to the back vowel counterpart. Also the tongue body is in a rather forward position because of the constriction which widens the sagittal distances at the larynx and the pharynx regions. These are the characteristics of the front vowels

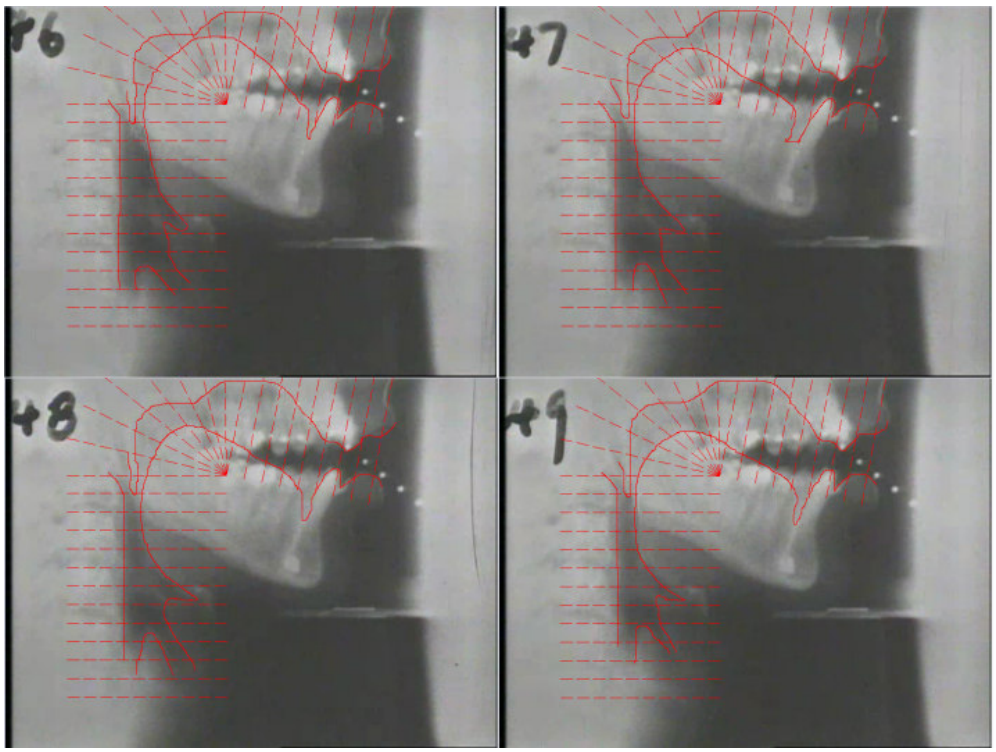

Figure 3-14 : X-ray images of /ga/ or /ka/ transition (left to right).

In addition to show the differences of back and front vowel cases, in fact these two figures are the demonstrations of the *co-articulation* property of the human articulatory system, which allows the human being to utter sounds at higher rates. This property is a simple adaptation of articulatory mechanism which uses common vocal tract movements and articulators to be able to reach to a high speech rate.

This property is one of the main ideas to derive a model for plosive to vowel transitions thus it will be used in the model as the natural utterances necessarily involve this property.

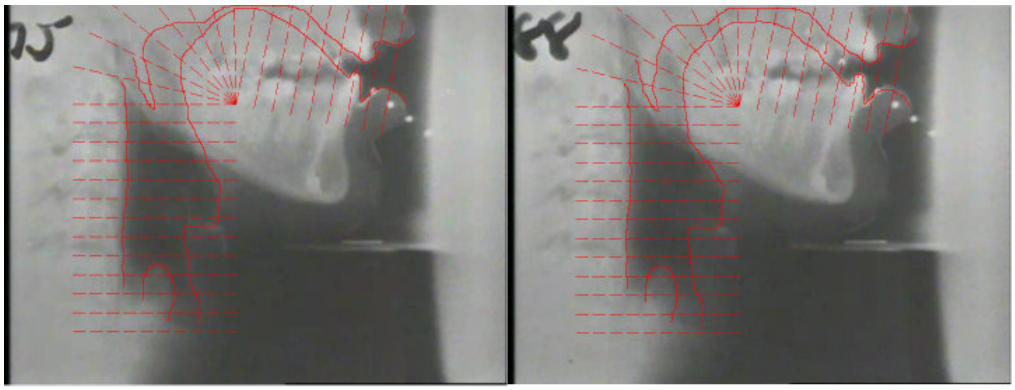

Figure 3-15 : X-ray images just at the beginning of the utterance /ki/ (left one) and at the end (right one).

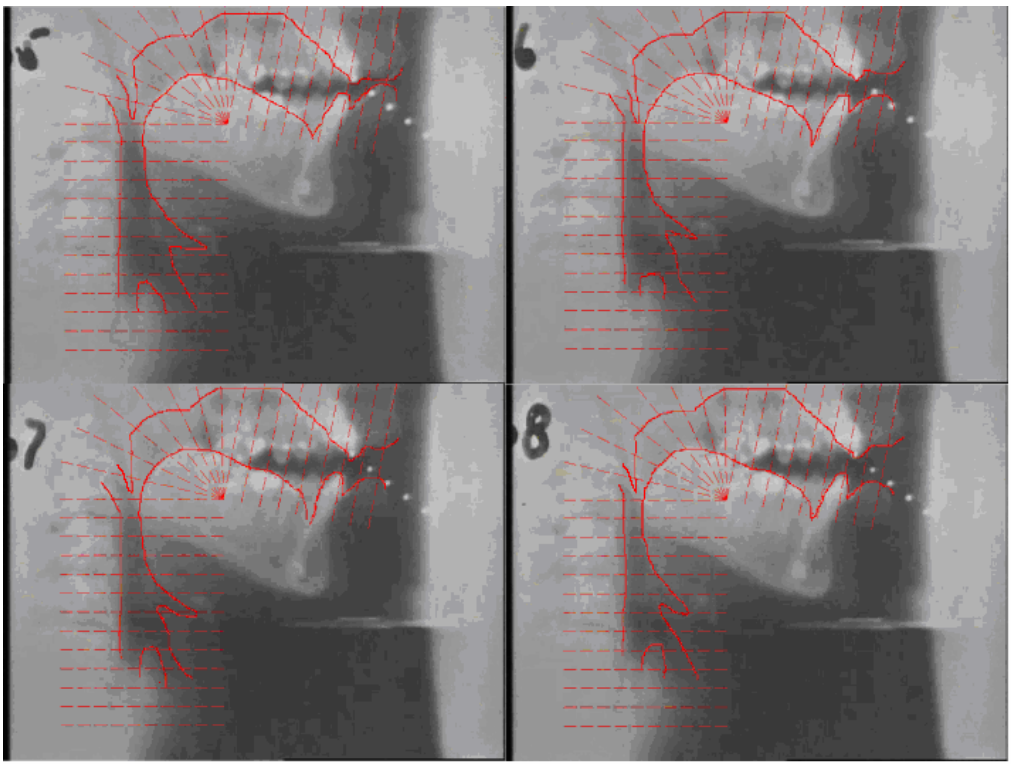

Figure 3-16 : X-ray images of /ba/ or /pa/ transition (left to right).

3.1.3.2. Labial Snapshots

In Figure 3-16, labial plosive to back vowel transition is examined. Unlike the velar case, the labial transition seems to be more stationary. The reason is explained (as the type of the constriction) in the previous sections. At the beginning of the snapshots, the vocal tract shape seems to be very close to the upcoming vowel except for the lips and teeth openings. Progressing from the labial stop to the back vowel, the jaw moves a little bit downwards with the more lowered tongue tip, widening the narrow openings at the labial and the alveolar parts, resulting a vocal tract shape for a back vowel. There are not any other differences observed except the above mentioned ones. It may be interesting to compare the positions of the articulators just before the articulation of /pa/ and /pe/ examining Figure 3-17 and Figure 3-16. At the left side of the figure the vocal tract is about the utter the labial plosive with a following front vowel /e/. Compared to the back

vowel case (See first snapshot of Figure 3-17), the tongue seems to be in a more forward position. Besides, the tongue body is in a higher position for /e/ utterance. On the other hand for back vowel, the root of the tongue is in a back position compared to the front vowel which is again the result of the co articulation property of the vocal tract. To sum up, every sagittal distance is different from each other except for the constriction for two cases.

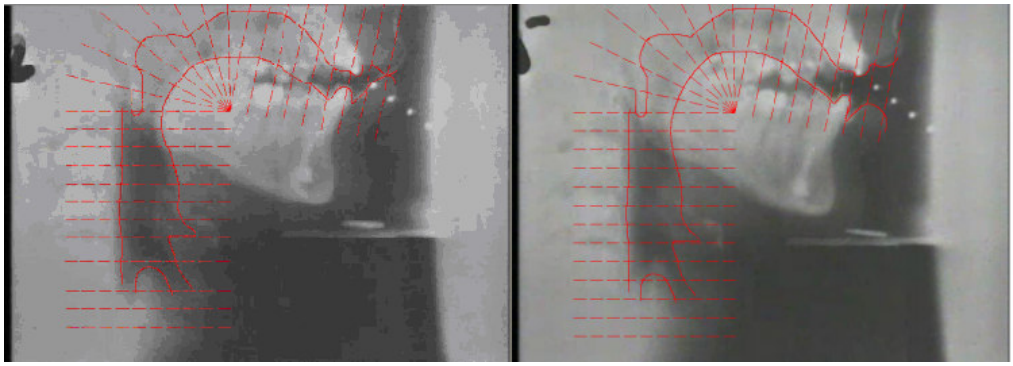

Figure 3-17 : X-ray images just at the beginning of the utterance /pe/ (left one) and at the end (right one).

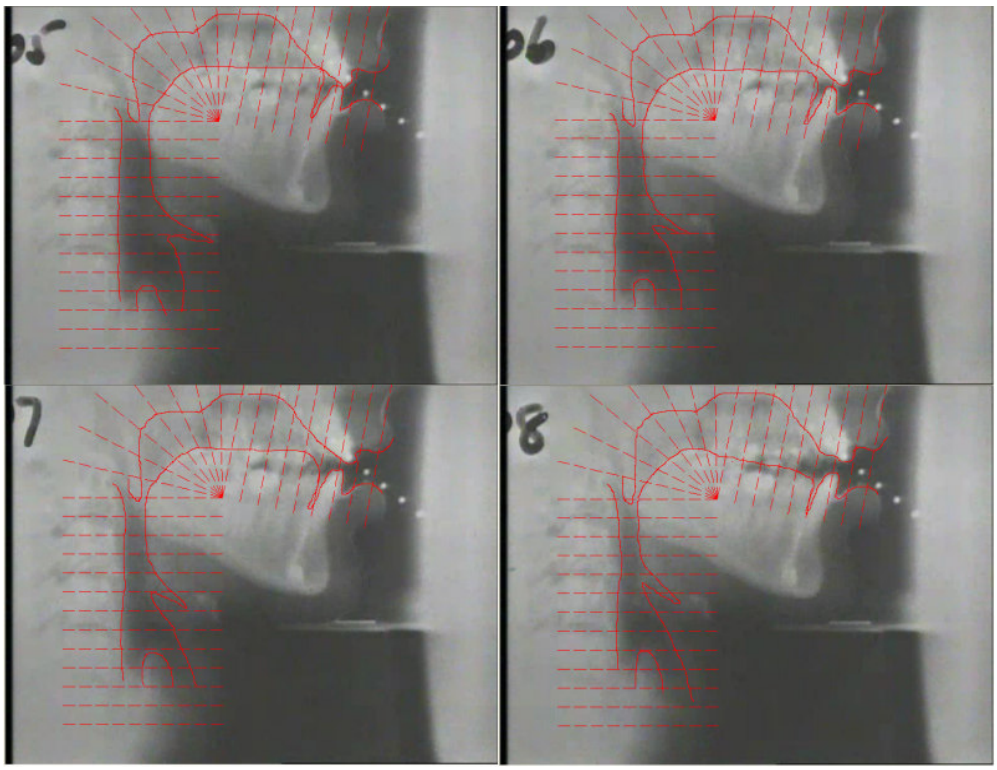

Figure 3-18 : X-ray images of /da/ or /ta/ transition (left to right).

3.1.3.3. Alveolar Snapshots

The last transition is an alveolar plosive to back vowel transition as illustrated in Figure 3-18. Examining slides, it is observed that the tip and the blade of the tongue constrict the vocal tract at the front hard plate before teeth and the tongue body is lowered as time goes by. At the end, the tongue is totally lowered and withdrawn towards pharynx to utter the upcoming back vowel. The comparison of front and back vowel transitions can be done examining Figure 3-19 that left image is for ti / di and the right one is for ta / da transitions. The difference of the two transitions can be easily observed so that the back of the tongue is more raised for the front vowel and the body is closer to the hard and soft palates. Therefore the pharynx region is wider for front vowel as expected.

To sum up the empiric observations done by the help of the X-ray movies, we can say that the plosive to vowel transitions are greatly in the dominance of the

following vowel. The dominance can be accepted as a result of the co-articulation property which is common for all types of plosive vowel transitions. The co-articulation does not only govern the shapes of the vocal tract for plosives, but also decide on the constriction place for velar stops. But it is certain that, this adaptation greatly increases the utterance rates while decreasing the complexity of the vocal tract movements.

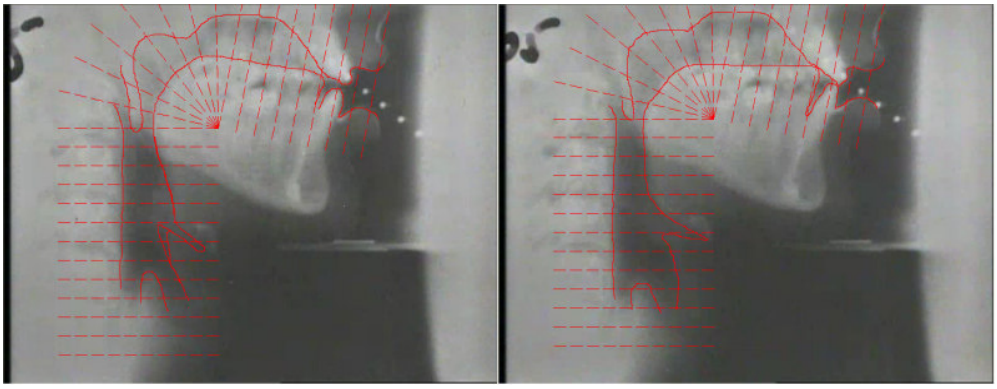

Figure 3-19 : The moments the speaker is about to utter /ti/ (left) and /ta/ (right).

3.1.3.4. Estimation of Vocal Tract Area Functions Making Use of Sagittal Distances

After empirical observations, it is essential to examine the vocal tract movements analytically for a detailed characterization of the process. For this purpose, we have taken the first 133 milliseconds (4 frames) of each transition and segmented the vocal tract into 22 to 24 pieces depending on the vocal tract length, in each frame following the same path. For each frame, the sagittal distances of each segment are recorded from the pixel data. The pixel data will be translated to the area data by using some estimation techniques. As a simplest technique cylindrical tube assumption is used i.e. the areas are calculated by simple circle area equation $A = \pi r^2 = \pi d^2 / 4$ where r is the radius and d is the sagittal distance measured. In this study, to estimate the areas we have used a more sophisticated method, explained in the following sections, to get more accurate

results. After the estimation of the cross sectional areas, the reflection coefficients can be calculated by Equation A.13. The procedure is repeated for all three kinds of plosives. At the end, the reflection coefficients and the area plots are analyzed and discussed for the transition model.

Area Computation

X-ray vocal tract images are very useful for analysis of vocal tract positions for speech production. But unfortunately this method is not as powerful as MRI (Magnetic Resonance Imaging) when the exact areas of the vocal tract are interested. This is because the X-ray can visualize the vocal tract only within the planes that slices the human body vertically whereas the MRI can visualize both vertical and horizontal slices. Thus MRI method can extract the exact area values of the vocal tract. But unfortunately it is hard to obtain MRI data of the vocal tract because of the limited studies in this field and it is very costly to obtain personally. In addition to the high cost, MRI takes too long to visualize the stationary vocal tract whereas X-ray can give us sagittal distances in real time while the speaker is talking. Therefore, for the analysis of the transitions, we can make use of available X-ray videos instead of the MRI although the measurements of the X-ray data will not represent the real areas of the segments.

In the literature there are some techniques to estimate vocal tract areas from mid-sagittal distances measured from the cross-sectional images. But before estimating the areas, the mid-sagittal distances are needed to be measured correctly. The X-ray images have 480×360 resolution and they are actually the concatenated and scaled versions of the real X-ray films. A practical approach to measure the length of each pixel can be done by the help of the approximate vocal tract length calculated from the MRI data. According to Story and Hoffman [25] the vocal tract is approximately 174 millimeters long when it is articulated to produce the vowel /a/. The sound data is extracted from the video data and a snapshot that is the moment for a natural /a/ is obtained. Then slicing the vocal tract for small lengths and averaging the experiment over large number of ensembles, a pixel is found to be a square of one side is 0.34 millimeters long.

The vocal tract is a hook shaped space that lies from glottis to lips. Because of the curvatures it is hard to decide on the sagittal distances. So another problem

is to measure the distances with the correct angles with respect to the vocal tract image. There are lots of segmentation techniques proposed in speech literature like Maeda [26], Mermelstein [7] etc. which generally slice the vocal tract in three regions. The labial, alveolar and hard palatal regions are sliced almost vertically, the laryngeal region is sliced horizontally and pharyngeal region, soft palate and velum are sliced with inclined lines as if they come from a common center point located at the centroid of the tongue body. Having common properties with others, following configuration is proposed by Mohammed [6] for his studies.

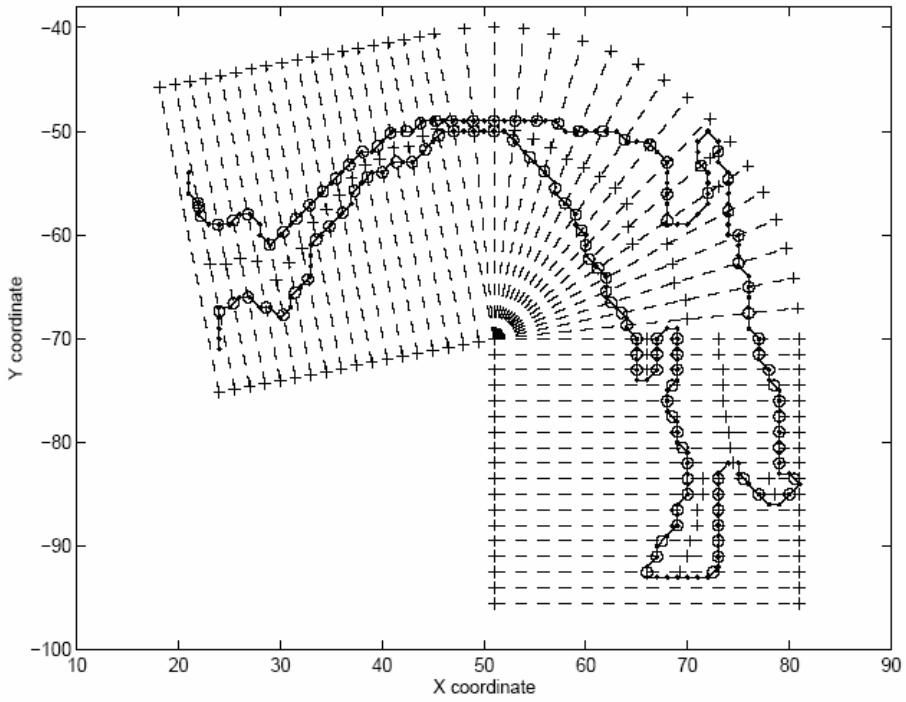

Figure 3-20 : The guide for slicing the vocal tract.

In Mohammed's configuration, the vocal tract is sliced horizontally starting from the glottis end. The horizontal slices continue with inclined lines in pharyngeal region sweeping an angle of $9\pi/16$ radians covering also the velar region. The remaining parts are sliced with parallel lines making an angle of $9\pi/16 - \pi/2$ radians with vertical. The circle arc is sliced in a fashion so that the arc distances

at the semi radius of the curvature are chosen to be equal to the distance between two parallel adjacent lines in horizontal and vertical regions so that the average length of the vocal tract is tried to be sliced with equal length of portions.

All X-ray snapshots are sliced according to the guide grid as proposed by Mohammed [6]. The (consequent) images are tried to be sliced with the stationary reference grids to obtain healthy measurements for transitions. After slicing the vocal tract, the last step is to estimate the area functions from sagittal distances. However, this is not a simple procedure which has a general method to follow. In the literature there exist many studies struggling with this problem. Heinz and Stevens [29], Maeda [26], Baer et al. [16] etc proposed methods with a general relation $A = \alpha\, d^{\beta}$ alternative to πr^2 where A is the tube area, d is the sagittal distance measured and α, β are the specified constants. But none of these methods serves a general solution to the problem because they are all results of studies based upon few casts, cadavers or subjects which make these methods precisely applicable to the information they use. In fact, the ambiguity is a natural result of the morphologic differences from person to person. As an example, Figure 3-21 shows two vocal tract area functions for vowel /a/ which is uttered by two male native speakers of English having the same weight but differing eight centimeters in height [16].

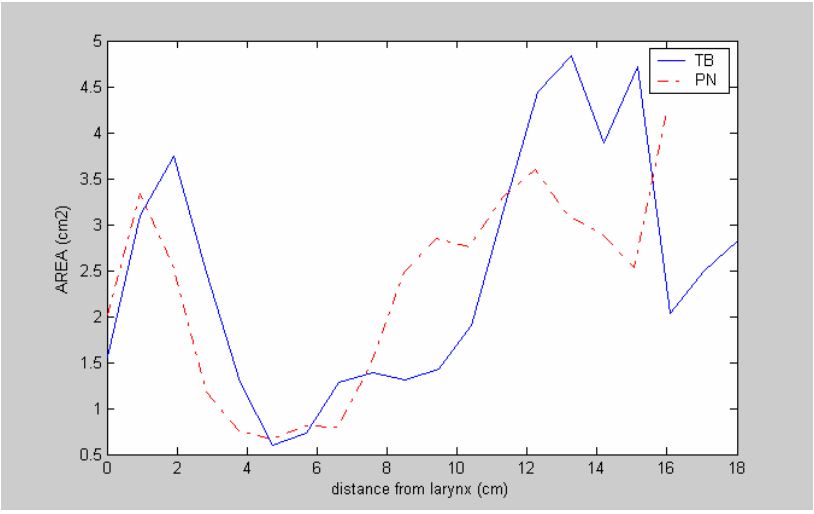

Figure 3-21 : The vocal tract area functions of two male subjects obtained by MRI.

The morphologic difference of the subjects TB and PN from the vocal tract side of view is clear from the Figure 3-21. However main characteristics of the vocal tract for vowel /a/ can be accepted to be similar. The characteristics of other vowels uttered by different subjects also seem to be similar. With these motivations and the absence of opportunities which supply continuous vocal tract area functions in time domain (such as continuous MRI), it will be reasonable to use vocal tract area function estimation techniques mentioned before as a guideline for our purposes.

αβ model was first proposed by Heinz and Stevens [29], according to his studies he accepted tongue as a flat surface then he calculated the cross sectional areas as a function of sagittal distances through general form. Then Sundenberg [30] tried to synthesize nine Swedish vowels from X-ray tracings. He divided the vocal tract in two regions: the buccal zone and the pharynx. With help of three male subjects he determined that α varied from 2.07 to 2.63 and β from 1.33 to 1.47.for the buccal zone. For the pharynx, which is the most difficult region to model in these kinds of studies, he proposed a non-monotonic relation between α and β with Fant's tomographic data [31]. However these calculations were contrary to the Heinz and Stevens's calculations for pharynx. Such studies one confuting some parts of others continued in the literature without an agreement in a common general method. Table 3-4 shows the results of three such studies.

If the table is analyzed even in a rough manner, the variety of the coefficients calculated by different studies greatly discourage us to use them for a consisted study. However following observations can be done to encourage the study of conversion:

1. Plosives are uttered by constrictions. The vocal tract gets nearly the shape of the following vowel before the utterance as a result of co-articulation property. Thus, a major change is not expected except for the constriction region during the transition (See X-ray images). This property decreases the complexity of the conversion, as minor sagittal changes will lead minor area changes in the transition. Therefore the losses will be minimized for the plosive to vowel transitions as many of the tubes are expected to be stationary during the transition.

2. In addition to partial stationary assumption of the transition, we are interested in normalized areas of the vocal tract sections. Therefore $\alpha\, d^{\beta}$ type of conversion reduces to d^{β} which will also reduce the complexity of the calculations.

Table 3-4 : The experimental results of alpha beta model proposed by different studies

	Martin et al. [15]		**Alain et al. [13]**		**Perrier et al. [11]**	
Zone	**α**	**β**	**α**	**β**	**α**	**β**
Larynx floor	1.563	0.040	1.437	1.088	1.36	1.5
Larynx	1.515	1.690	1.437	1.088	2.74	1.5
Low pharynx	1.824	1.381	2.138	1.036	3.04	1.5
Mid pharynx	1.347	1.626	2.138	1.036	2.37	1.5
Oro pharynx	0.731	1.806	2.138	1.036	2.37	1.5
Velum	1.393	1.079	1.813	0.903	2.89	1.5
Hard palate	1.334	1.514	1.221	1.273	2.91	1.5
Alveolar	1.875	1.193	2.946	0.191	2.12	1.5
Labial	4.697	2.481	2.540	0.578	2.12	1.5

Examining the table, the ambiguity can easily be observed. But with our assumptions, the alpha columns will not be considered in the modeling procedure. Focusing into betas, there are large differences between them since their values

are in the interval [1-2] except for the larynx and the labial regions. The labial regions will be examined in more detail with continuous camera snapshots and the pharynx region will be accepted to be changed properly making use of MRI data. Thus if the sagittal distances are calculated with beta model, beta changing from 1 to 2, we will have reasonable area functions to analyze. Therefore, for the three transitions (velar, alveolar and labial) β approximation is done for each tube. Beta is changed between 1 and 2 with 0.1 incremental steps. The approximated areas are normalized for each beta value and for each tube. The normalized exponential tracks with suitable time constants are also calculated for each approximation. All approximated data are plotted for each tube for every transition type in the following figures.

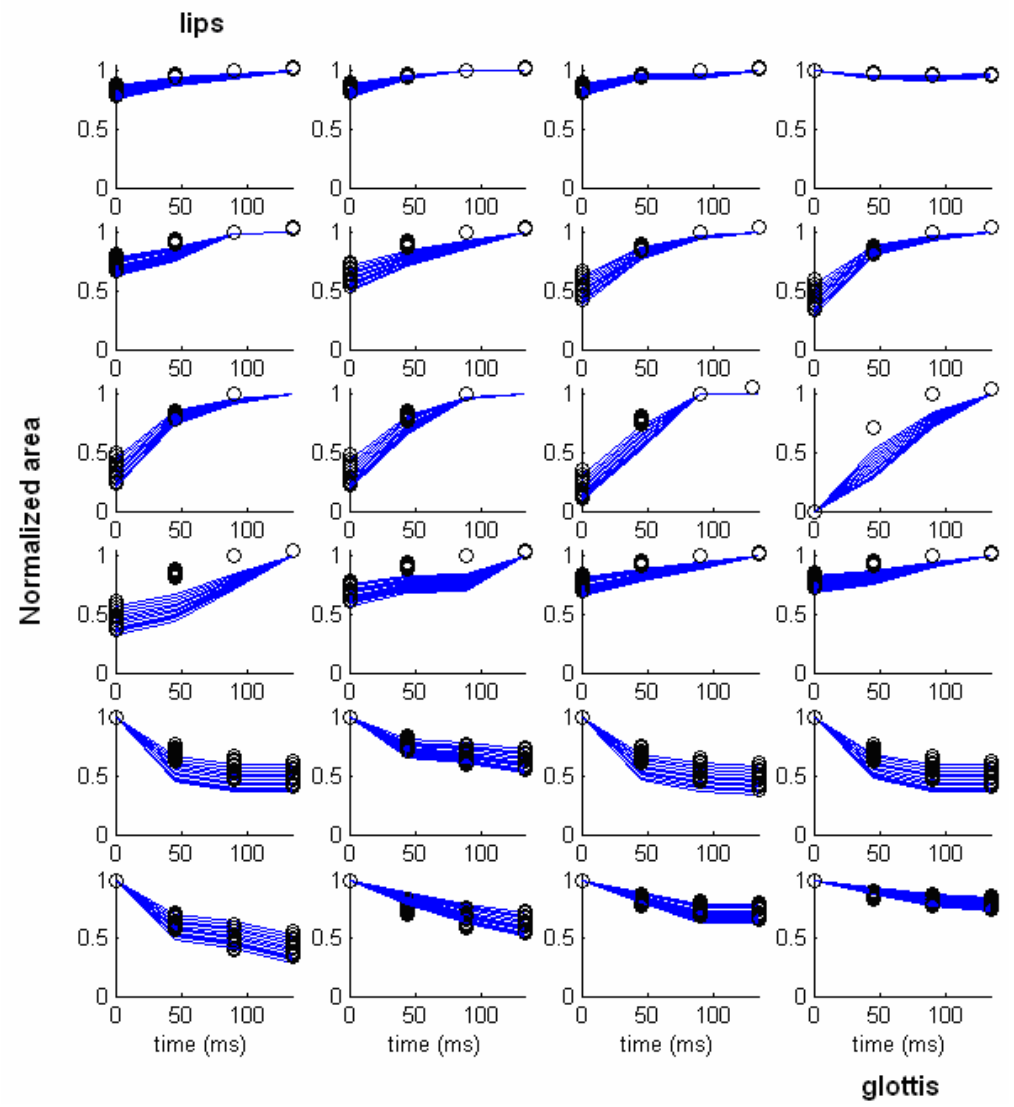

Figure 3-22 : The estimated area trajectories of each tube for /ga/ transition with different beta values (solid lines, smallest values for beta = 1) and fitted exponential tracks for each estimated area values (dots). (Left to right, first one is the lip area).

Figure 3-22 is for velar plosive to back vowel transition. The estimations show that the velar region is enlarged mainly whereas the pharyngeal and larynx regions are straitened. Labial and alveolar regions are not affected except for small enlargements in the hard palate. This was indeed expected because the velar constriction done by the tongue body was pulled back to form the back vowel. Therefore these regions were evacuated by the tongue body whereas the pharyngeal region and larynx were filled with the tongue root. Rather than the verbal description of the transition the quantities or the characteristics of the quantities are more important. Therefore approximated and normalized areas for all beta values are superimposed with exponential tracks for each tube with a single time constant. The track is represented with dots whereas the estimated areas are shown by solid lines. All powers of sagittal distances seem to be exponential like, that's why all estimates also seem to be exponential. This can be explained in such a way that, any power of an exponential is also an exponential. This property will be illustrated theoretically in the following section. The exponential area changes are also observable for alveolar plosive to back vowel transition in Figure 3-23.

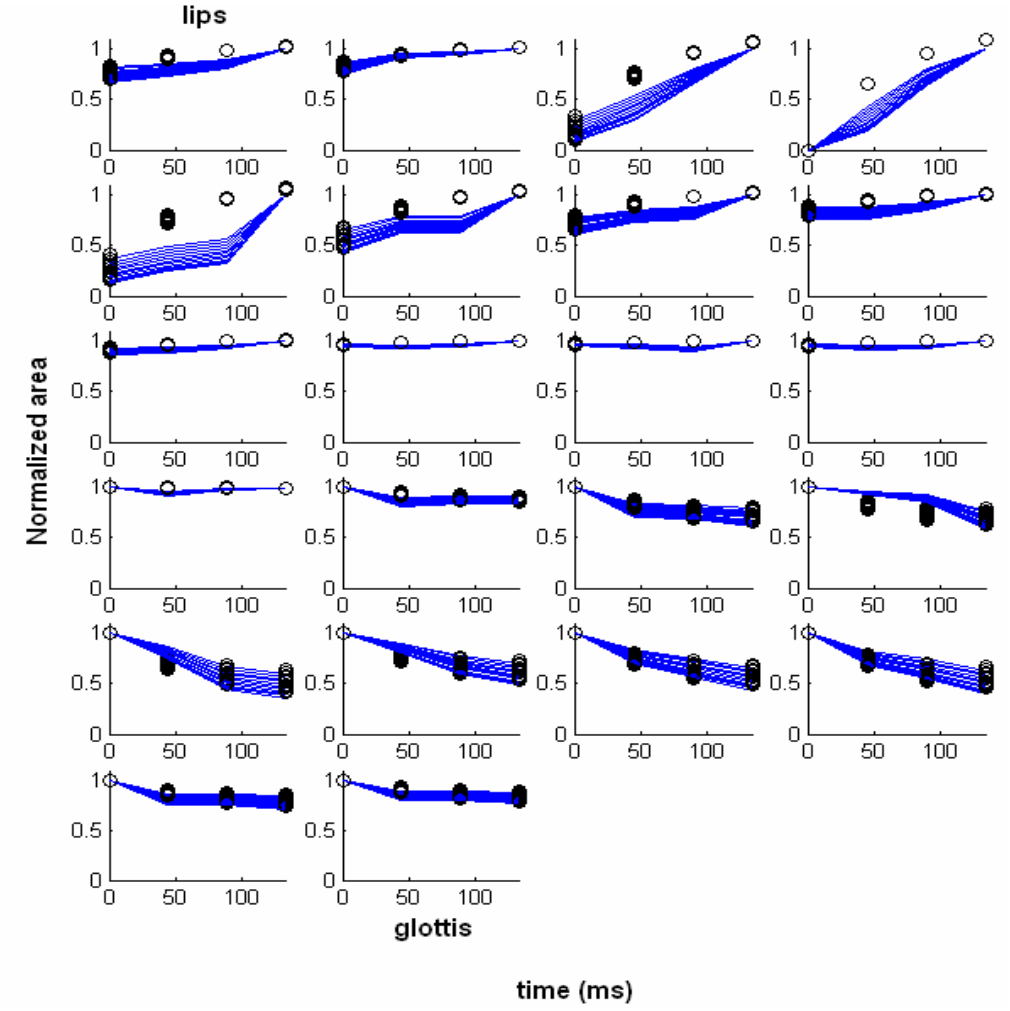

Figure 3-23 : The estimated area trajectories of each tube for /da/ transition with different beta values (solid lines, smallest values for beta = 1) and fitted exponential tracks for each estimated area values (dots) (Left to right, first one is the lip area).

Again back release straitens the pharyngeal region and larynx whereas the labial region is enlarged because of the relaxation at the hard palate. For this transition the co-articulation property is more observable than velar case because it is seen that the labial and velar regions are not affected by the transition with smooth stable estimated area plots. The last transition of back vowels, labial plosive to back vowel seems to be the most stationary transition among the three (See Figure 3-24). This is because the tongue does not move during the transition, so does not cause any major area changes. The major change during the transition is the downward movement of chin thus the downward movement of the jaw and the lips enlarging the alveolar opening. The movement of the chin straitens the pharyngeal and larynx region that's what all can be seen from the graph. In addition to the back vowel case, we will examine one more transition with a front vowel for the sake of completeness. Figure 3-25 shows one of these transitions. Comparing the transitions with previous ones, the plosive to open vowel versions of the transitions are observed to be more stationary because except for labial plosives, the constriction is done by the front movement of the tongue body which is withdrawn for back vowels whereas it is hold stationary for a front vowel utterance.

As a summary, after X-ray investigations, the vocal tract areas for the plosive to vowel transitions seem to be exponentially changing. These areas are estimated with alpha beta model which is commonly used in speech literature. The evaluation of each tube in its group provided the simplification of removing the alphas. For betas the areas are estimated when beta changed from 1 to 2 which is a commonly used range in related studies. The estimated areas with betas also seemed to be exponentially changing. The main reason of this is observed to be the exponential changing sagittal distances. Among the three types of the plosives, the labial ones are observed to be the most stationary transition whereas from the vowel side of view front vowels are observed to be more compatible to oral constriction because of the narrow structure of the vocal tract. As last words for this section, the exponential model seems to be reasonable for the area transition but still not knowing the beta values for each tube causes an ambiguity of the estimated time constant for each tube. But it may be reasonable to approximate a common time constant if the figures are analyzed one by one and

see that the extra beta values for each area do not cause marginal area changes thus the time constants are not greatly manipulated.

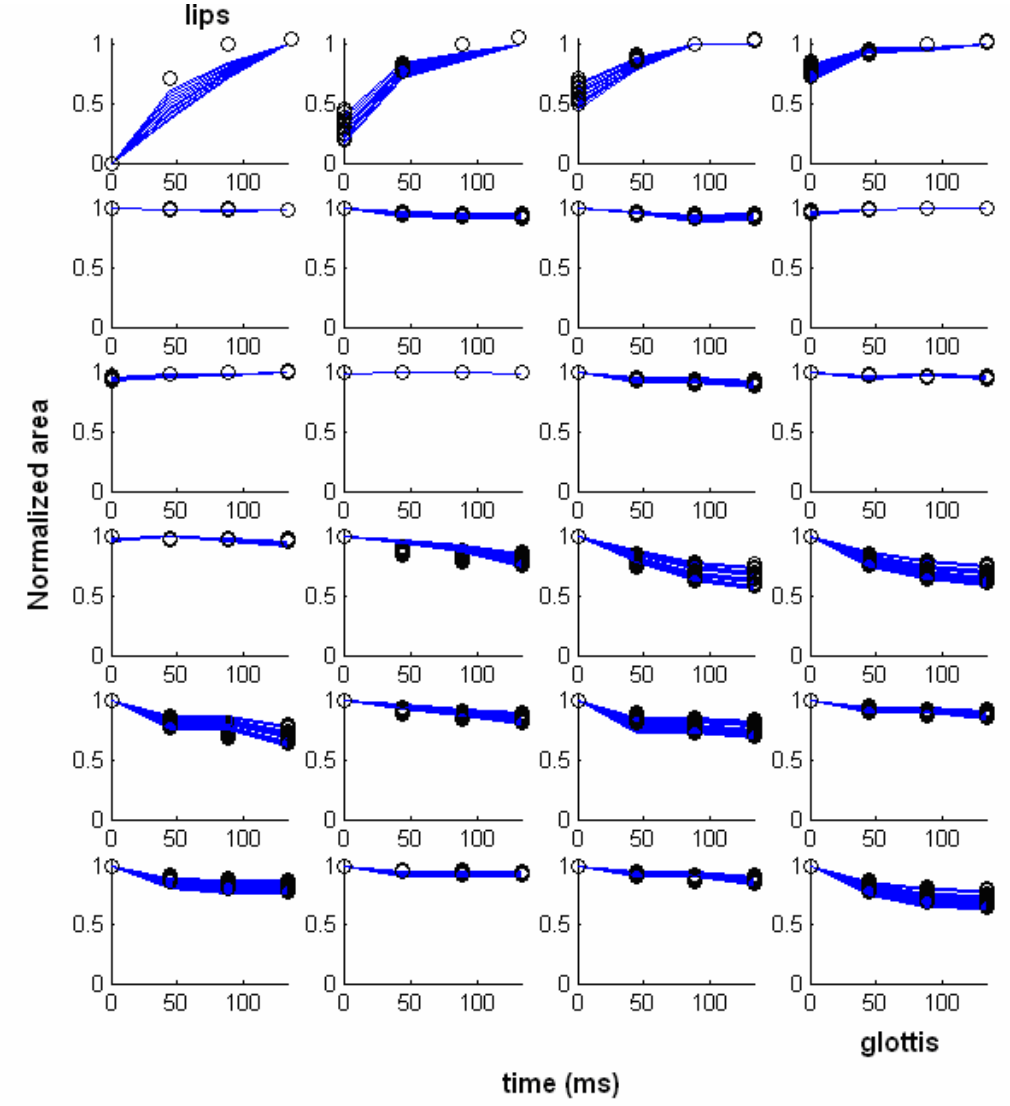

Figure 3-24 : The estimated area trajectories of each tube for /ba/ transition with different beta values (solid lines, smallest values for beta = 1) and fitted exponential tracks for each estimated area values (dots) (Left to right, first one is the lip area).

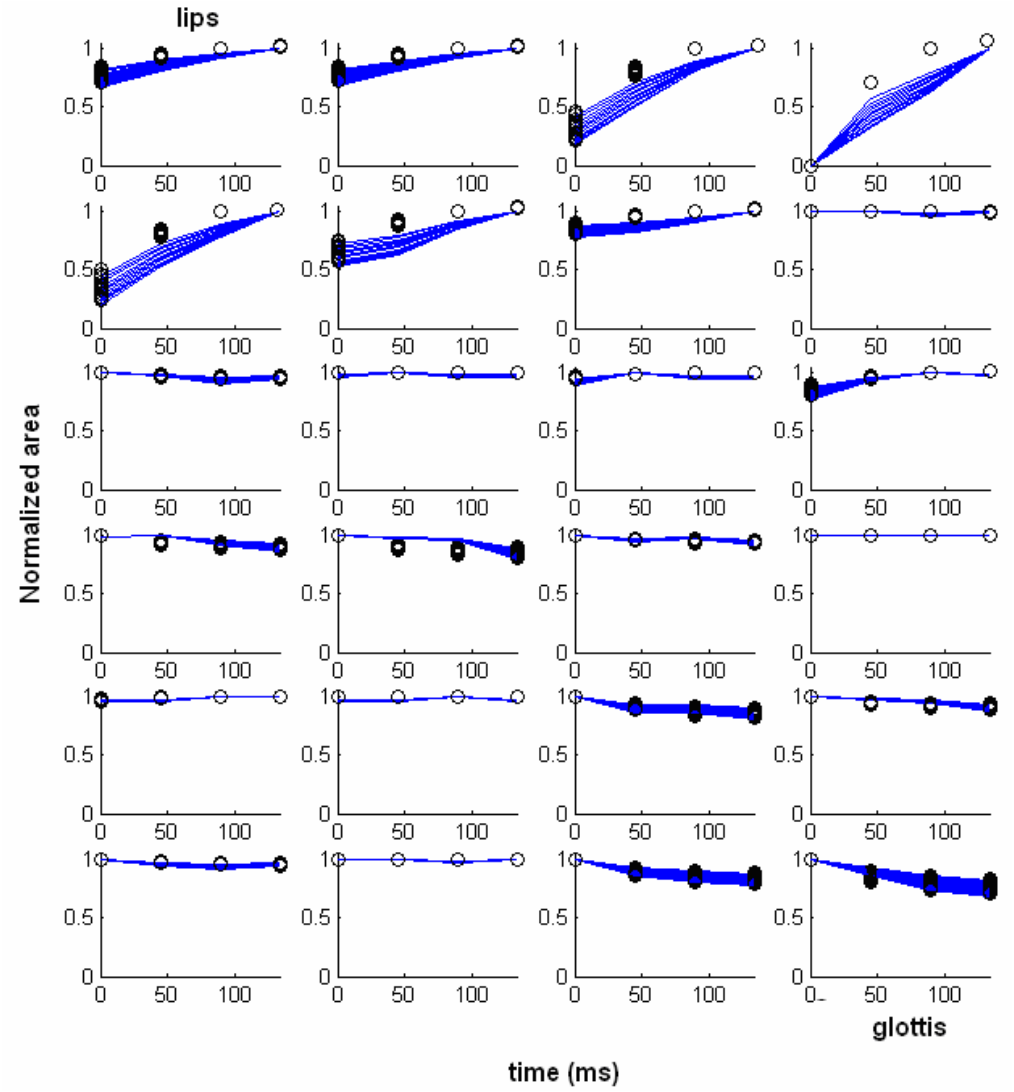

Figure 3-25 : The estimated area trajectories of each tube for /de/ transition with different beta values (solid lines, smallest values for beta = 1) and fitted exponential tracks for each estimated area values (dots) (Left to right, first one is the lip area).

Theoretical Illustration

Sagittal distances are observed to be changing exponentially. Assume that a distance, as a function of the time can be approximated as an exponential function in the form:

$$d(t) \cong \delta + (\mu - \delta) e^{-\gamma t} \qquad 0 \le \delta, \mu \le 1 \tag{3.3}$$

where μ and δ are the normalized initial and final distances and $1/\gamma$ is the time constant. According to the alpha beta model an area can be approximated by using the sagittal distance. Thus the area can be written in terms of the exponentially fitted sagittal distance as follows:

$$A(t) \cong \left(\delta + (\mu - \delta) e^{-\gamma t}\right)^{\beta} \tag{3.4}$$

In Equation 3.4, α is omitted to reduce complexity. A good approximation to this expression can be obtained as follows:

$$\frac{A(t)}{\delta^{\beta}} \cong \left(1 + \left(\frac{\mu}{\delta} - 1\right) e^{-\gamma t}\right)^{\beta} \tag{3.5}$$

For $\left(\frac{\mu}{\delta} - 1\right) e^{-\gamma t}$ is small compared to 1and considering approximation Equation 3.6

$$(1 + x)^{\beta} \cong (1 + \beta x) \;\; for \;\; small \; x \tag{3.6}$$

$$A \cong \delta^{\beta - 1} \left(\delta + \beta(\mu - \delta) e^{-\gamma t}\right) \tag{3.7}$$

Of course Equation 3.7 is a very rough estimation having bad performance especially around t = 0. However from that derivation we can see that the time constant of the sagittal distance is approximately equal to the time constant of the

estimated area function. On the other hand for a good approximation δ term should be approximately 0. Rewriting Equation 3.4 with δ = 0 and μ =1we get,

$$A \cong \left(e^{-\gamma t}\right)^{\beta} = e^{-\gamma \beta t} \quad \textbf{(3.8)}$$

Knowing that $1 \leq \beta \leq 2$, even in the extreme case (δ = 0) with the worst beta (β = 2) the estimated area function is observed to have the half of the time constant of the sagittal area function. Thus we can conclude that, having an exponential sagittal distance function, the area function is also exponential having a time constant comparable with the time constant of the distance function. Figure 3-26 and Table 3-5 show various curves and time constants for different beta, initial and final area configurations. Analyzing the data, the new time constant is observed to be between the half and twice of the initial time constant.

Table 3-5 : Fitted and approximated time constants for different beta , initial and final area configurations.

Constriction $\left\|\left(\frac{\mu}{\delta}-1\right)\right\|$	0.33		1		8		∞	
Beta values	1.5	2	1.5	2	1.5	2	1.5	2
Fitted τ	1.06τ	1.13τ	1.35τ	1.77τ	0.73τ	0.57τ	0.66τ	0.5τ
Approx. τ	τ	τ	τ	τ	τ	τ	τ	τ

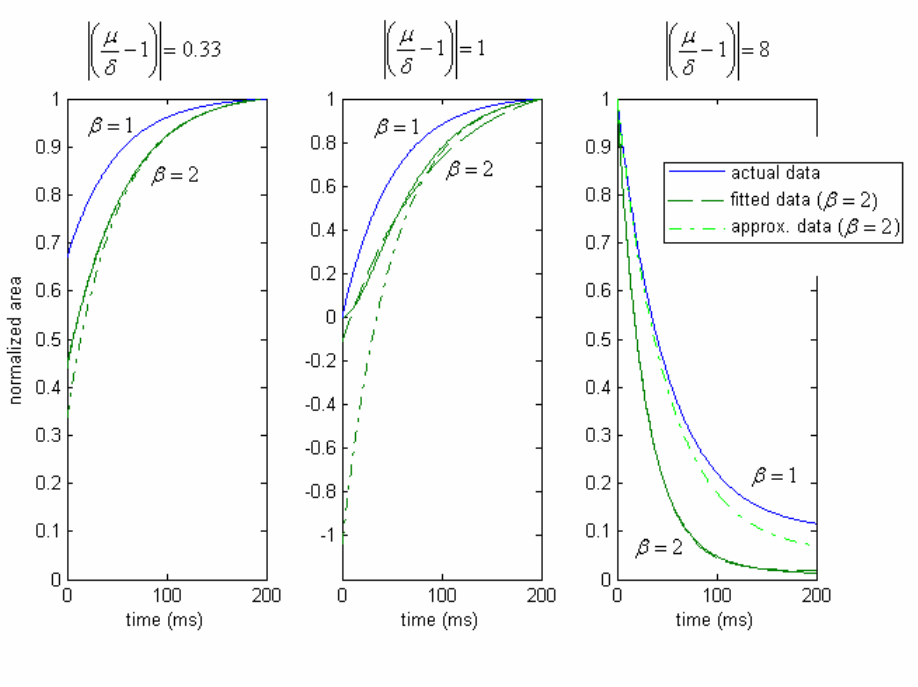

Figure 3-26 : Actual, fitted and approximated curves for different configurations.

3.1.4. Lips Video Examinations

To examine the releases better, fast frame rate labial snapshots directly taken from face will be useful. Figure 3-27 shows the configuration of camera setup. The camera used for the snapshots is Canon Powershot A700 and its capturing rate is sixty frames per second with 320 * 240 pixels resolution. During the capture, the face is forced to keep stationary with respect to the camera lens and the captured snapshots are transferred to the paint program for analysis. In the program the lips openings are tried to be modeled with ellipses according to Mermelstein [7]. The lip openings are projected into elliptical surfaces (as if a rouge wearing women kissing a piece of paper) and obtained elliptic like surface areas are tried to be fitted with different parameter ellipses In addition to the elliptical lip areas, the sagittal distance between the tip of the lips and upper teeth is tried to be estimated. The procedure is applied to five labial plosive to vowel utterances which are /ba/, /be/, /bi/, /bo/ and /bu/ and one alveolar plosive

utterance /da/ in both directions. For each transition, normalized sagittal distances and lip areas are plotted for analysis.

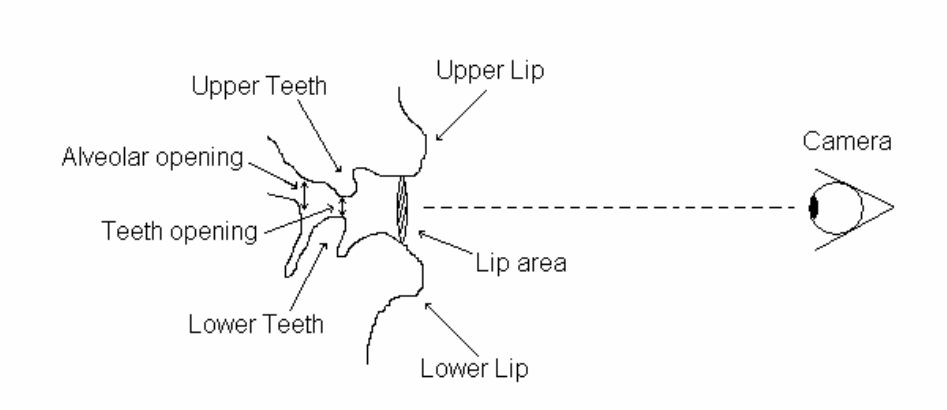

Figure 3-27 : The illustration of camera recording setup.

3.1.4.1. Labial Snapshots

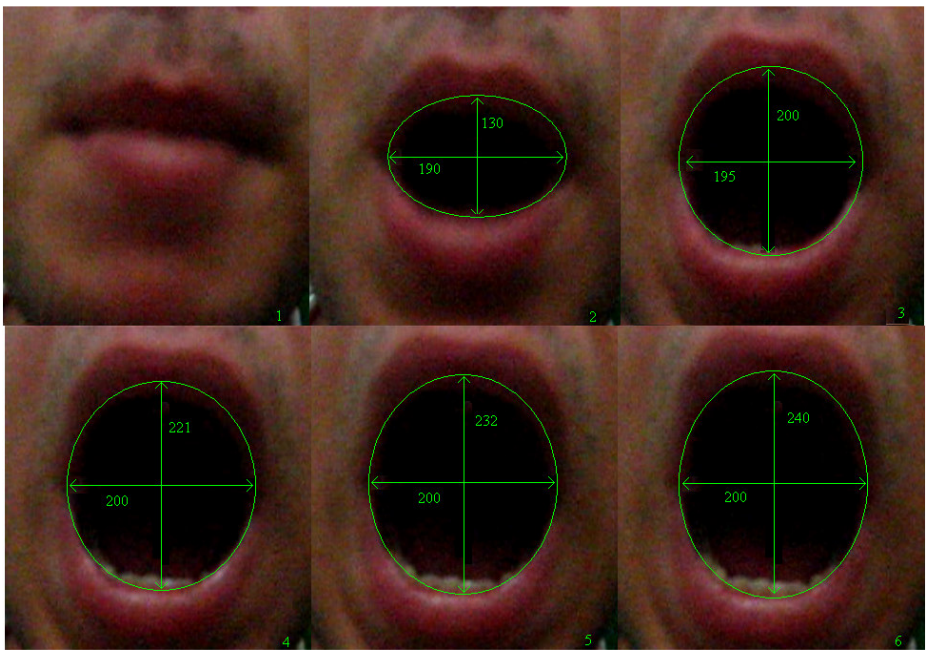

Figure 3-28 : First 100 ms of the front view of /ba/ transition.

For /ba/ utterance, the fitted ellipses are as in Figure 3-28. The total transition lasts nearly hundred milliseconds and more than fifty percent of the lip opening is formed in the second snapshot with a decreasing rate of increase in the following snapshots. The summary of all kind of labial transitions in both directions is given in Figure 3-29. And from that figure it is easy to see the strong relationship between the sagittal distances and the areas of the lips openings for utterances /ba/, /be/ and /bo/. This is because the width of the mouth opening does not change for these transitions. On the other hand for /u/ and /i/ the width of the mouth opening also increases like the sagittal distance. The function of the width of the opening is also observed to be exponential for these two utterances but the time constant of the width function is quite smaller than the sagittal distance. Therefore the width converges to the limit quicker than the sagittal distance. Thus the relationship between sagittal distance and lip area for these two utterances are not much distorted and accepted to be exponential, like other three. Finally for consistency, it will be suitable to compare X-ray sagittal distances with the direct snapshots. Comparing tube 1 of Figure 3-24 and /ba/ transition in Figure 3-29, it is clear that both seem to be exponential with similar time constants. It is obvious that the time constant can change from person to person and it may also change for the same person depending on the existence of stress on the transition. The transition uttered with stress is expected to have smaller time constant compared to normal utterance.

In addition to the exponential movement of the articulators, the time constants of the movements and the formants are compared in Table 3-6. For labial case the time constants of the lip area and formants are calculated to be 22.4 ms and 20 ms respectively (in the sense of root mean square error). This interesting observation shows the correlation between the formants and the articulator movements supporting our claims about plosive to vowel utterances.

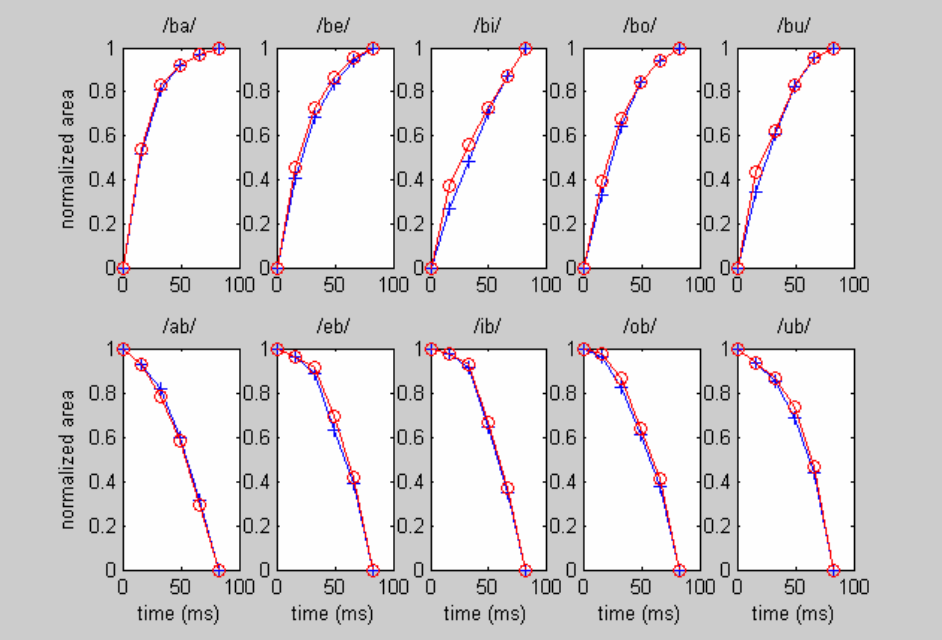

Figure 3-29 : Comparison of labial lip areas (bubbles) and sagittal distances (pluses) in both directions.

Figure 3-30 : First 100 ms of the front view of /ab/ transition. The lip openings are modeled with ellipses. The units are dimensionless.

3.1.4.1. Alveolar Snapshots

The lips video examination technique is also applied to alveolar transitions. The same setup is used as illustrated in Figure 3-27. The sagittal distances at teeth and lips are tried to be estimated. The sagittal distance at the tongue end is also tried to be estimated knowing that the upper teeth is stationary with respect to the body frame. In addition to the sagittal distances, the lip opening area is fitted with ellipses. All these parameters are plotted in Figure 3-33. These measurements are also consistent with the X-ray investigations once more. The exponential trajectory is encountered at the alveolar and labial regions of alveolar releases as in the case of X-ray distances (Notice that blur increases at last snapshots in Figure 3-31, and decreases in Figure 3-32 as the velocity increases).

In addition to the articulatory movements, the comparison of time constants of the articulatory movements and the formants may be interesting. Table 3-6 shows the time constants of the /da/ formants and the teeth to tongue distance. They are calculated to be 30 ms and 36 ms respectively (in the sense of root mean square error). Again a high correlation between the vocal tract movements and the formants which are actually the output of the articulator movements show that alveolar plosive to vowel transitions are also highly related with exponential articulations as labial transitions.

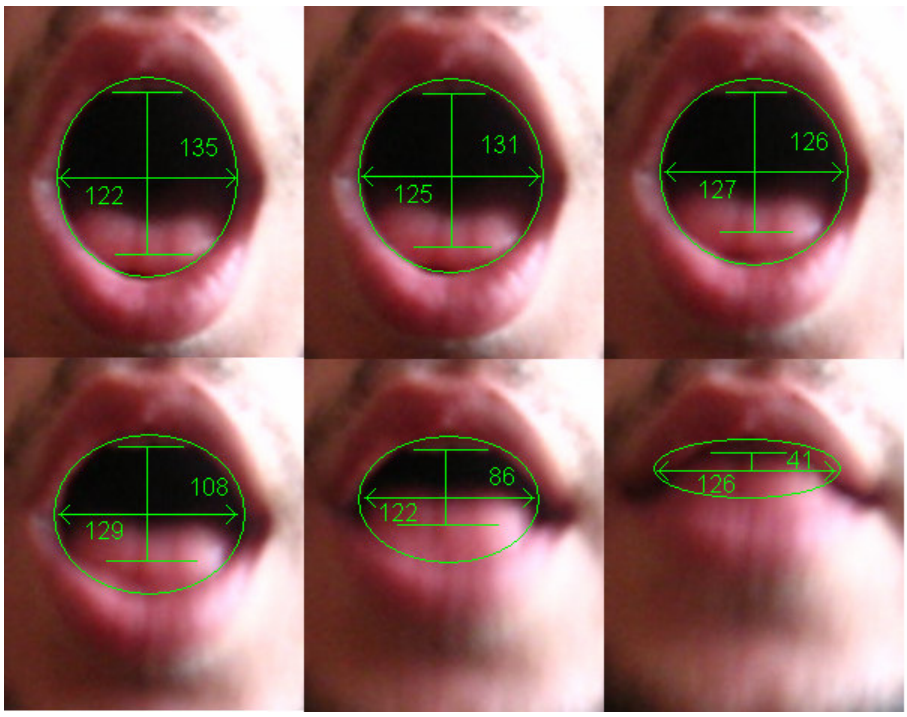

Figure 3-31 : First 100 ms of the front view of /ad/ transition. The lip openings are modeled with ellipses. The upper teeth to tongue tip distance is tracked with capital I symbol. The units are dimensionless.

To sum up, the video measurements of velar, alveolar and labial transitions showed that, the sagittal distances and the cross-sectional areas of visible parts of these utterances are exponential like functions. The measurements may seem to be useless for velar case but the lip area is the main factor for labial utterances and the importance of tongue blade movement is added to the lip opening for utterance of labial releases. The character of these movements are also observed to be exponential like functions and the cross-sectional areas are estimated to be exponential in the X-ray investigations, therefore the reliability of these estimations is improved and a second almost real data is found to be compared with the estimated ones.

In addition to the exponential utterance, formants and movements are observed to be correlated to each other by inspecting the time constants of the two for labial and alveolar transitions.

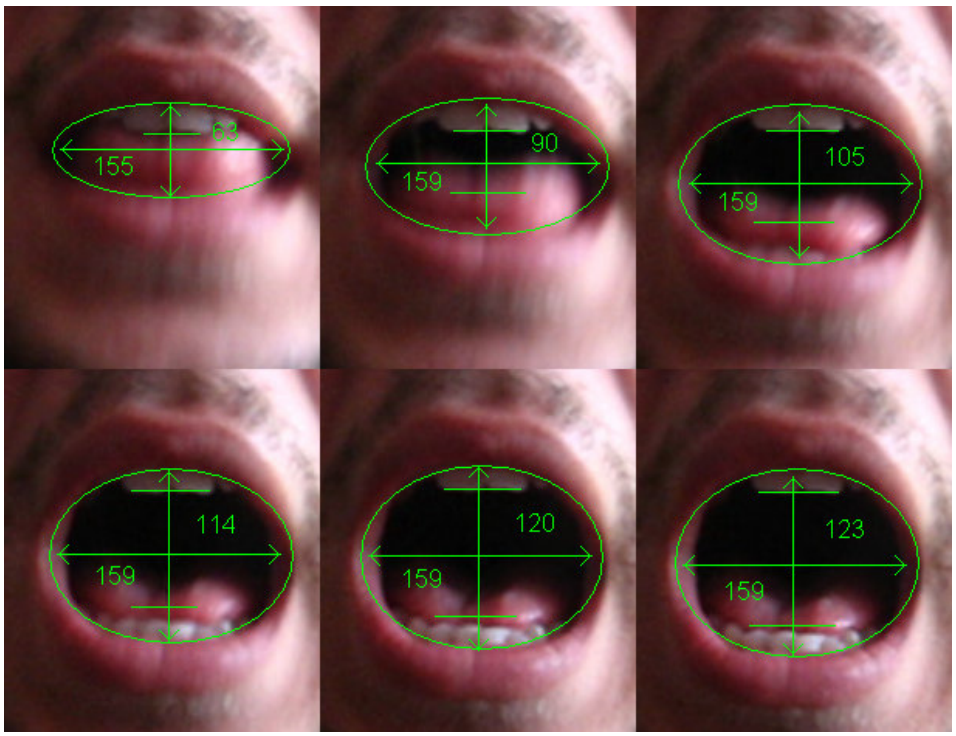

Figure 3-32 : First 100 ms of the front view of /da/ transition. The lip openings are modeled with ellipses. The upper teeth to tongue tip distance is tracked with capital I symbol. The units are dimensionless.

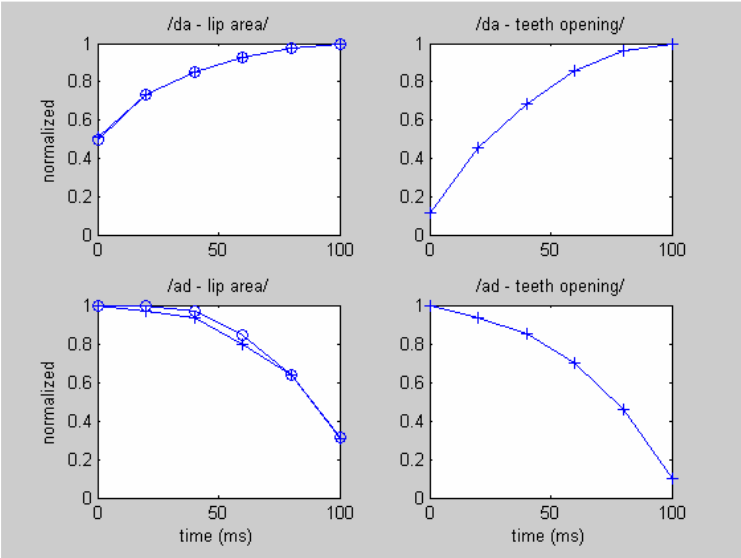

Figure 3-33 : Comparison of alveolar lip areas (bubbles) and sagittal distances (pluses) in both directions.

Table 3-6 : Time constants of different parameters

Variable	/ba/ formants	/ba/ lips area	/da/ formants	/da/ teeth to tongue distance
Time Const. (ms)	20.0	22.4	30	36

3.2. Fricative to Vowel Transitions

In previous sections labial, alveolar and velar unvoiced / voiced plosives are analyzed (/p/, /t/, /k/, /b/, /d/, /g/) with different aspects. It may be useful to compare the plosives with the fricatives of the same type, i.e. with the same place of articulation. The reason behind this comparison can simply be related to the acoustic similarities of fricatives and plosives. First of all, the articulation places of the fricatives are nearly same as the plosives, thus the formant structures are also expected to be similar. Secondly, the articulation is done by partial constriction of the vocal tract from oral cavity which is also parallel with the plosives and finally, the existence of a matching couple of a fricative for all types of plosives shows the strong relationship between these two groups of consonants. Moreover in some languages' histories it is found that, the plosives are evolved from the fricatives of the same language. This research also reveals the morphologic common properties of these sounds [17] [25].

In Turkish, /z/ and /s/ are alveolar; /v/ and /f/ are labial fricatives of voiced and unvoiced types respectively. Unfortunately velar fricatives do not exist in Turkish, that's why it may be useful to get the missing couples from English for analysis purposes. Completing the set with velar, alveolar and labial fricatives of voiced and unvoiced types, six vowel-fricative-vowel transitions are recorded. The preceding and succeeding vowels are chosen to be /a/ to compare the results with previous parts easily and to keep consistency. Following spectrograms are obtained to be analyzed.

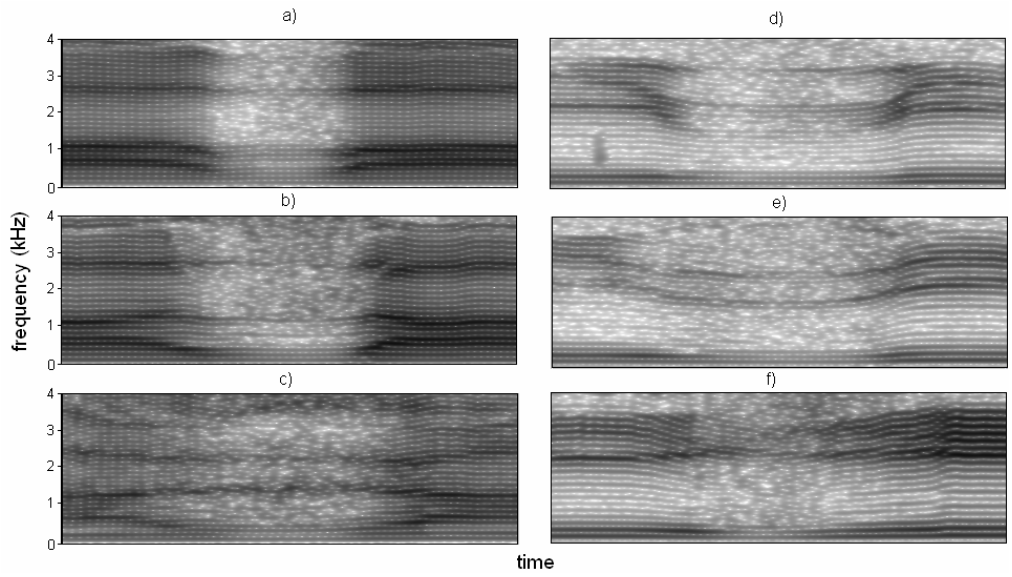

Figure 3-34 : Spectrograms of the utterances of (a) /a v a/, (b) /i v i /, (c) / a z a /, (d) / i z i/, (e) /a ğ a/, and (f) /i ğ i/.

Fricatives have significant energy for all time intervals compared to their plosive counterparts. This property is preserved in both directions which can be related to the partial closure of the vocal tract. Small openings at the constriction places provide the leakage of the propagating sound waves to radiate from the mouth. Thus, the main difference of fricatives from the plosives is the continuity of the sound waves if the voice bar for voiced plosives is not accepted to be a radiated sound wave. In frequency domain the formant tracks of the fricatives seem to be similar those of plosives (Compare Figure 3-1 and Figure 3-34). The similarity is violated at the partial closure part as an expected result because the formants of the fricatives are not inhibited with the total closure. Remembering the plosive theory, the total constriction of the vocal tract has the low pass characteristics inhibiting the large frequencies. Unlike plosives, the fricatives have stronger high frequency components because of that reason. In addition to the filtering effect, decrease of the supra glottal constriction increases the high frequency components of the excitation [20] having more noisy components at the output. This observation can be the second major difference between two groups. But despite the differences, the smoothness and the behavior of the formants are strong enough to keep the similarities of these sounds. The main reason for these

similarities can be related to the vocal tract movements, thus it may be wise to inspect X-ray images for these two sounds.

Figure 3-36 shows X-ray snapshots of a alveolar fricative /z/ to vowel /e/ transition. It is hard to find a major difference between Figure 3-37 and which is also an alveolar plosive to vowel /e/ transition (See also Figure 3-38). The only difference is the partial constricted vocal tract that is clearly observable from the images. Formants and the X-ray images show lots of similarities (See Figure 3-35 also) but are these similarities effective over the perception of these two sounds?

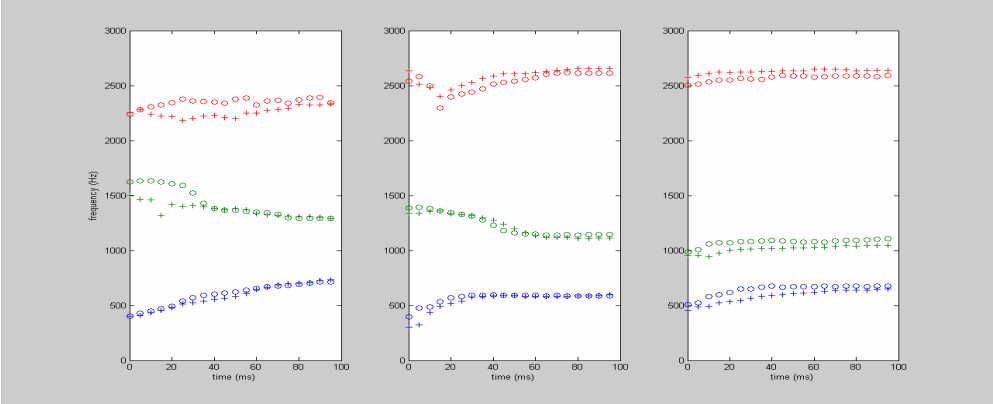

Figure 3-35 : Formants of the plosive transitions (bubbles) and inhibited fricative transitions (pluses) for velar, alveolar and labial cases.

As a curiosity, the fricative vowel transitions are clipped from proper places of the time waveforms such that the formant transitions are nearly same with the plosive to vowel transitions. The modified sound samples are asked to 10 subjects to be recognized. Following table is constructed for examining the recognition rates.

Table 3-7 : Clipped fricatives and recognition rates as the corresponding plosives

Utterance	/va/	/fa/	/za/	/sa/
Percept As	/ba/	/pa/	/da/	/ta/
Rate	%100	%100	%100	%100

It is obvious to see that, all kinds of fricative vowel transitions are percept as their plosive counterparts if the proper modifications are done. These modifications simply account for the inhibition of the sound waves by vocal tract constriction which is done manually by clipping the waveform. The inhibited formants of the transitions are plotted with the super imposed plosive trajectories in Figure 3-35 which actually shows the reason of high recognition rates.

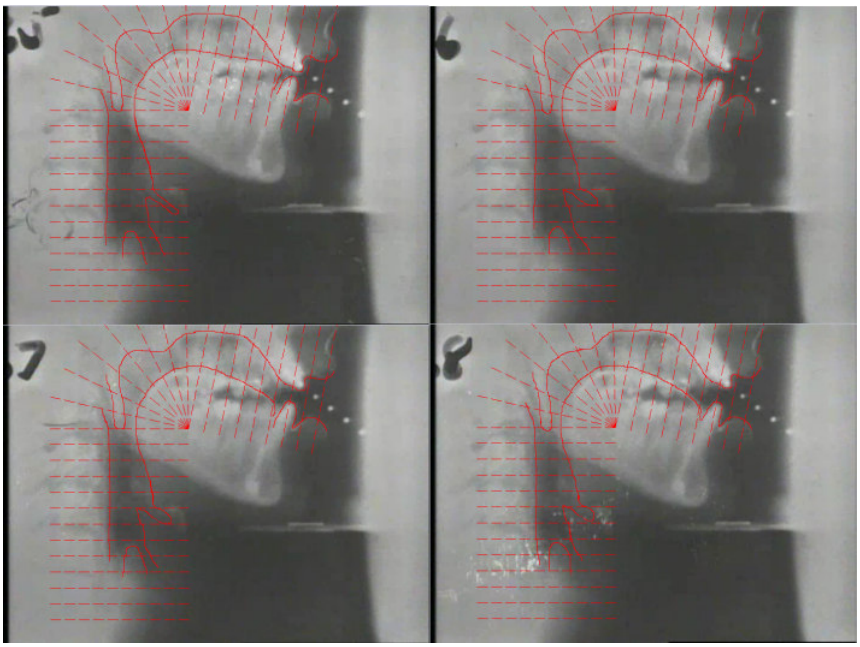

Figure 3-36 : X-ray images of /ze/ transition (left to right).

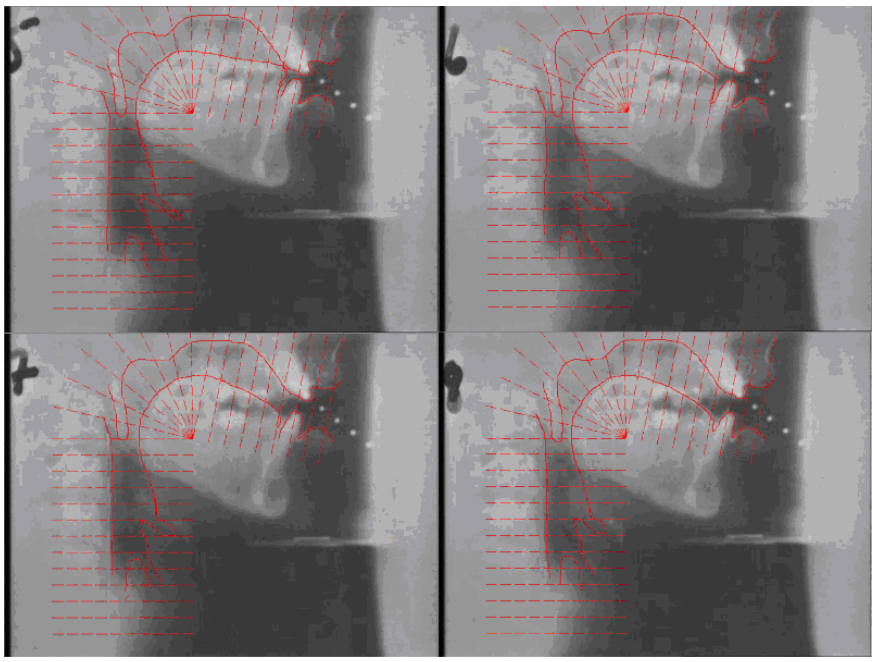

Figure 3-37 : X-ray images of /de/ transition (left to right).

As a last study, the camera snapshots of /za/ transition are taken for better alveolar and labial analysis. 133 ms. Interval starting from the beginning of the utterance of /z/ is plotted in Figure 3-39. The sound wave is extracted from the video stream using Virtual Dub software [32]. The sound wave is found to be percept as /da/ from the third snapshot (including). In addition to the snapshots, the normalized lip opening area and lip sagittal distance is plotted in Figure 3-40 (left). The teeth opening which is the same as alveolar sagittal distance in this case is plotted in the same figure (right). For analysis purposes, the regions of perception of /da/ are hatched with inclined planes.

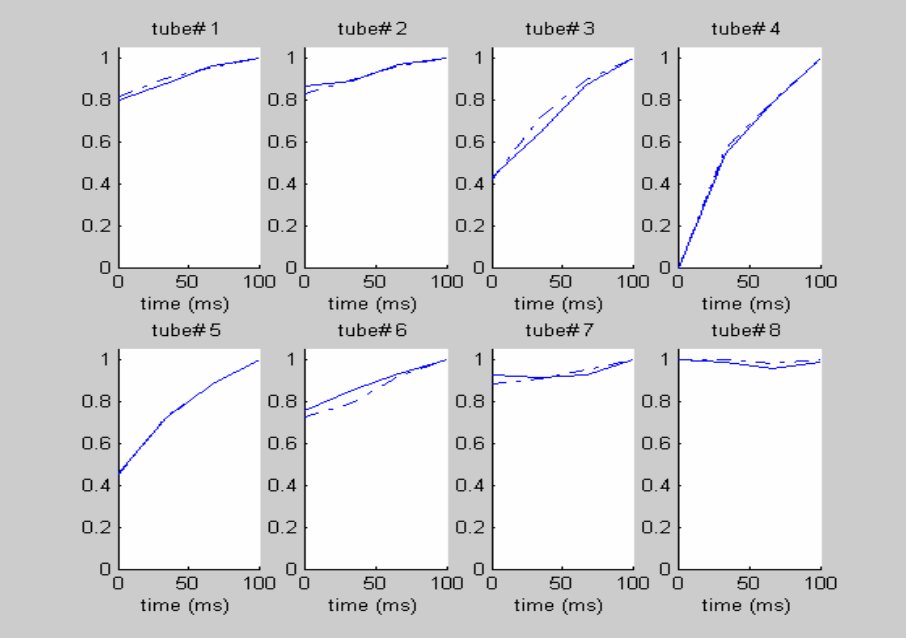

Figure 3-38 : Comparison of first eight tubes (from the lip end) of /de/ (solid lines) and /se/ (dashed lines).

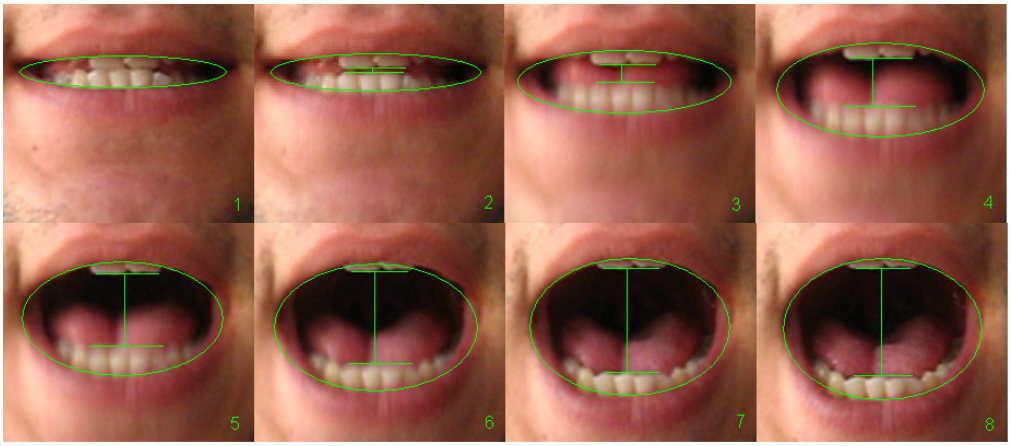

Figure 3-39 : First 133 ms of the front view of /za/ transition. The lip openings are modeled with ellipses. The upper teeth to tongue tip distance is tracked with capital I symbol.

As expected, /da/ region is observed to be exponential like and the time constant of the teeth opening and lip area are calculated to be very close to each

other if these three curves are fitted with exponential estimates in the sense of root mean square error.

Also comparing estimated lip area and teeth to tongue distance functions in Table 3-8, it is observed that time constants of these variables seem to be correlated to each other. Another comparison can be done among the time constants of /da/ and /za/ teeth to lip distances (excluding first two values of /za/). These values also seem to be correlated as expected.

Summing up, plosives and fricatives have lots of similarities in time and frequency domain. These similarities can be related to the place of articulation and vocal tract movements except for partial closure for fricatives. The reasons of the main differences in the spectrogram are strongly related to the partial closure and supra-glottal pressure level. Thus, plosives can be thought of the inhibited versions of fricatives which are revealed by the perception experiment. The inhibition of the vocal tract and the movements, thus formant changes, can be thought as the main factor of the production of plosives rather than the complex burst parts which give the name to that kind of sounds.

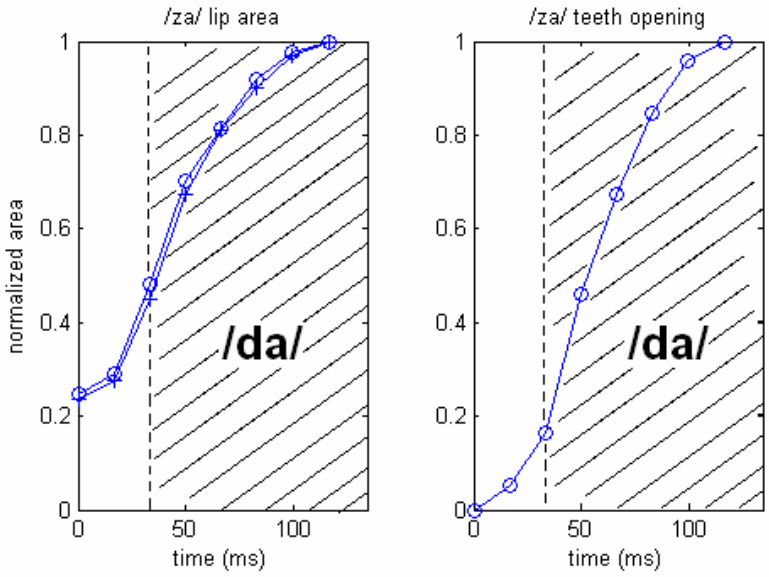

Figure 3-40 : Normalized lip area (o) superimposed with normalized lip opening (+) (left), Normalized teeth opening (right).

Table 3-8 : Comparison of (estimated) /da/ and /za/ time constants

Variable	/za/ lip area	/za/ teeth to tongue distance	/da/ teeth to tongue distance
Time Const. (ms)	38	44	36

3.3. Chapter Summary

As a summary of that section we can say that plosives are mainly uttered by exponential vocal tract movements in which the transition is mainly governed by the following vowel. In addition, the reverse transition is also observed to involve similar vocal tract movements.

On the other that, the exponential spectral changes are observed to be important from the perceptional side of view, this property is clearly illustrated with fricative to vowel transitions.

Lastly, fricatives are observed to be very similar to the plosives. Several examinations show that many properties of the two sounds are nearly the same.

CHAPTER 4

MODEL FOR PLOSIVE TO VOWEL TRANSITIONS

4.1. Summary of the Characterization

In previous chapters, detailed information about plosive to vowel transitions is given. In this chapter this information will be used to form a voiced plosive to vowel transition model. It would be better to give a short summary of previous chapters before proceeding to the modeling.

1. Plosives are simply formed by relaxing the vocal tract from the constriction region with a simultaneous excitation that starts from the glottis. For voiced ones, a reasonable assumption can be made about the simultaneous voiced excitation with the release. However this is not the case for unvoiced plosives.
2. The transition region for voiced plosives is voiced and the pitch periodic is observed to be continuous during the transition.
3. The formants changes of the transition are observed to be exponential like curves for voiced plosives. On the other hand formant structure of the unvoiced counterparts is also observed to be similar to the voiced ones. This is related to the similar vocal tract movements.
4. The perception mechanism of plosives is observed to be highly related with the formant difference between the beginning of the plosive and steady state vowel. The first formant is observed to be ineffective on the perception, on the other hand the second formant and in some cases the third formats are observed to be the dominant factors over the perception. For perception VOT region is observed to be important

for velar stops because of the rapidly converged second and third formants.

5. The co-articulation of vocal tract is almost always observed from the X-ray snapshots. Front vowels are observed to be more compatible to oral constriction. Hence the co-articulation is better realized.
6. As a result of co-articulation property, the areas except for the constriction regions are observed to be stationary during the transitions (especially for front vowels by the reason given in 5). Among plosives labial ones are observed to be the most stationary. Thus, a labial plosive to front vowel transition is found to be the most stationary transition.
7. In addition to 6, the areas are observed to be exponentially changing. Estimated time constants of different regions of the vocal tract are found to be different as expected.
8. Plosive to vowel transitions are observed to be the degenerated versions of fricative to vowel transitions which are uttered by partial constrictions. Inhibition of the partial closure region manually, yielded a plosive perception of the same kind. Besides, the format trajectories, X-ray snapshots and lips video examinations also showed that these two kind of transition show a great similarity.
9. Time constants of the formant trajectories and vocal tract movements are observed to be similar. This similarity is also observed if same type of fricative and plosives are compared.

4.2. Estimation of Reflection Coefficients

The most important factor of the transition i.e. the vocal tract movements is studied in the third chapter. It is observed that the vocal tract areas for each tube are changing by exponential like functions. It may also be interesting to analyze the reflection coefficients for a different perspective. The reason behind it lies in the truth of the dependence of acoustic model to the reflection coefficients rather than the cross-sectional areas. Remembering the tube model, the resonances of the filter were described by reflection coefficients of the vocal tract model.

Therefore, the estimated cross sectional areas of the utterances are converted to reflection coefficients. For the conversion, the relationship given in Appendix (Equation A.13) will be used. But unfortunately this equation requires area values instead of the sagittal distances. An estimation of areas is done in previous chapter that uses alpha beta model. Note that each tube is analyzed separately and alphas are cancelled with each other that remains betas for estimation. Also note that alpha beta model is an alternative to πr^2 representation of the area. A reflection coefficient can be written in terms of two adjacent tubes areas. Using alpha model we can write:

$$r_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} = \frac{\alpha_{i+1} d_{i+1}^{B_{i+1}} - \alpha_i d_i^{B_i}}{\alpha_{i+1} d_{i+1}^{B_{i+1}} + \alpha_i d_i^{B_i}} \quad \textbf{(4.1)}$$

As a property of the model, the alpha and the beta values are formed in groups with respect to the characteristic regions of the vocal tract (velum, pharynx etc. See also Table 3-4). Thus for two adjacent tubes in the same region, the Equation 4.1 can be simplified as:

$$r_i = \frac{d_{i+1}^{B} - d_i^{B}}{d_{i+1}^{B} + d_i^{B}} \quad \textbf{(4.2)}$$

Simplified equation may be suitable for similar adjacent regions in the same group, however for boundary calculations Equation 4.1 should be reconsidered.

First of all, the vocal tract area data is a continuous function typically varies from 0 to 8 cm^2. The alpha beta approximation slices this continuous function and fits group of alpha beta sets for the best representation. The number of groups is about 7 to 9 (See Table 3-4). In our model the vocal tract is sliced into 22-25 sections depending on the length of it. Therefore a reasonable approximation can model the boundary tubes in a different category as if they are in the same group because of the continuity of vocal tract and large number of tube samples that holds this continuity (In other words with 25 sections, the groups will not be strict so that the borders of the boundaries will not be well defined). So according to the

assumptions, it may be reasonable to use Equation 4.2 for all of tubes in the model. Using Equation A.13, the reflection coefficients of all tubes are calculated for all utterances and plotted with respect to time axis in Figure 4-1, Figure 4-2 and Figure 4-3. The reflection coefficients are calculated twice for $\beta = 1$ and $\beta = 2$ and they are plotted in same plots for comparison purposes.

In all figures, the types of the plosives can be easily estimated. The marginal changes in the reflection coefficients occur at the constriction places. These changes, like the area changes seem to be exponential but again with different time constants. As a result of co-articulation, surprisingly other reflection coefficients do not change much and many of them stay at the same value during the transition. The stationarity is especially observed at labial releases as the area functions are also stationary. Unlike, the velar releases show higher changes in the reflection coefficients. As a rough assumption it can be assumed that the reflection coefficients do not change much except for the constriction region for plosive to vowel transitions. For better understanding it will be good to examine reflection coefficients that are calculated by MRI area data that is obtained from utterances of the same person.

Figure 4-4 shows three of them. The first one is a velar plosive back vowel transition whereas others are alveolar and labial plosive front vowel transitions.

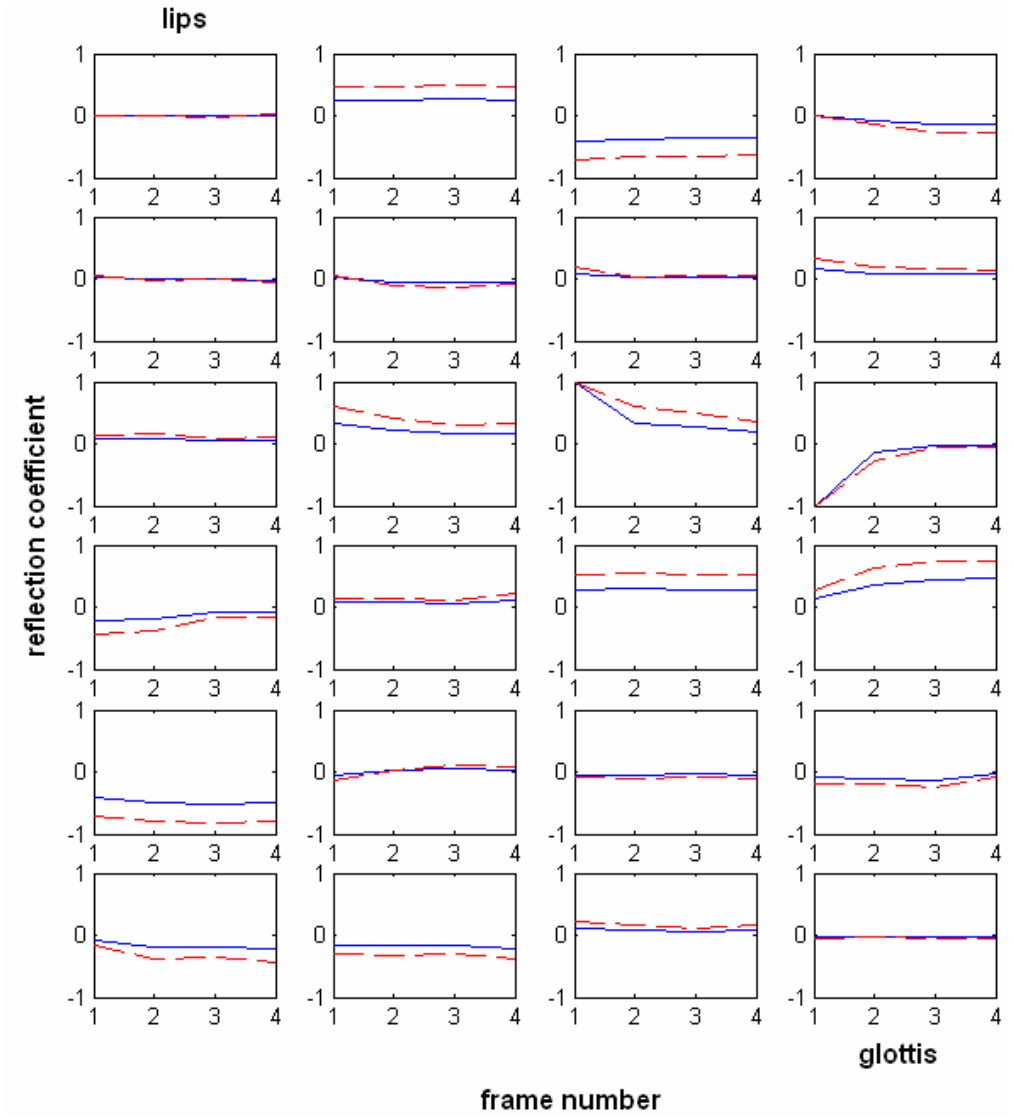

Figure 4-1 : The estimated reflection coefficients of estimated areas for /ga/ utterance with beta equals to 1 (solid lines) and beta equals to 2 (dashed lines).

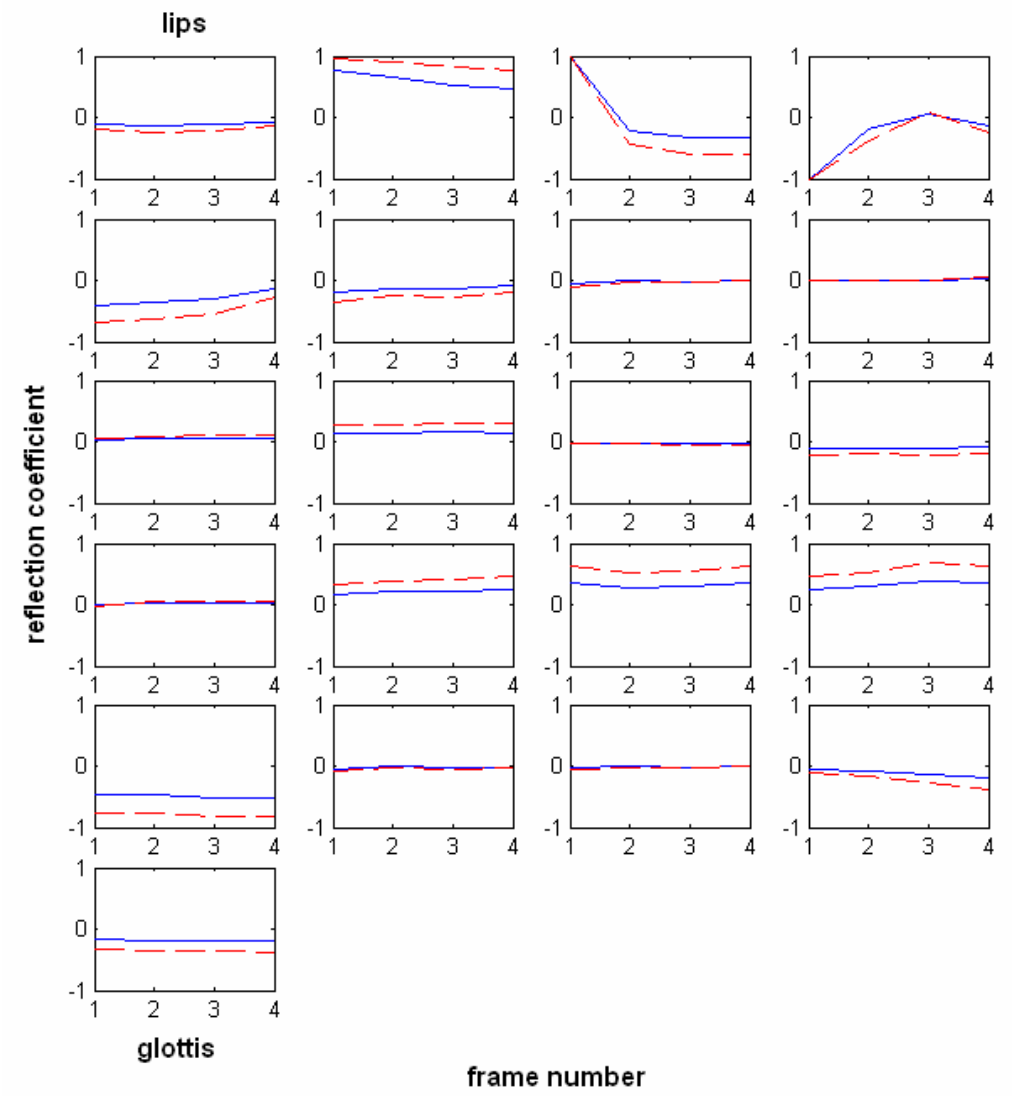

Figure 4-2 : The estimated reflection coefficients of estimated areas for /da/ utterance with beta equals to 1 (solid lines) and betas equal to 2 (dashed lines).

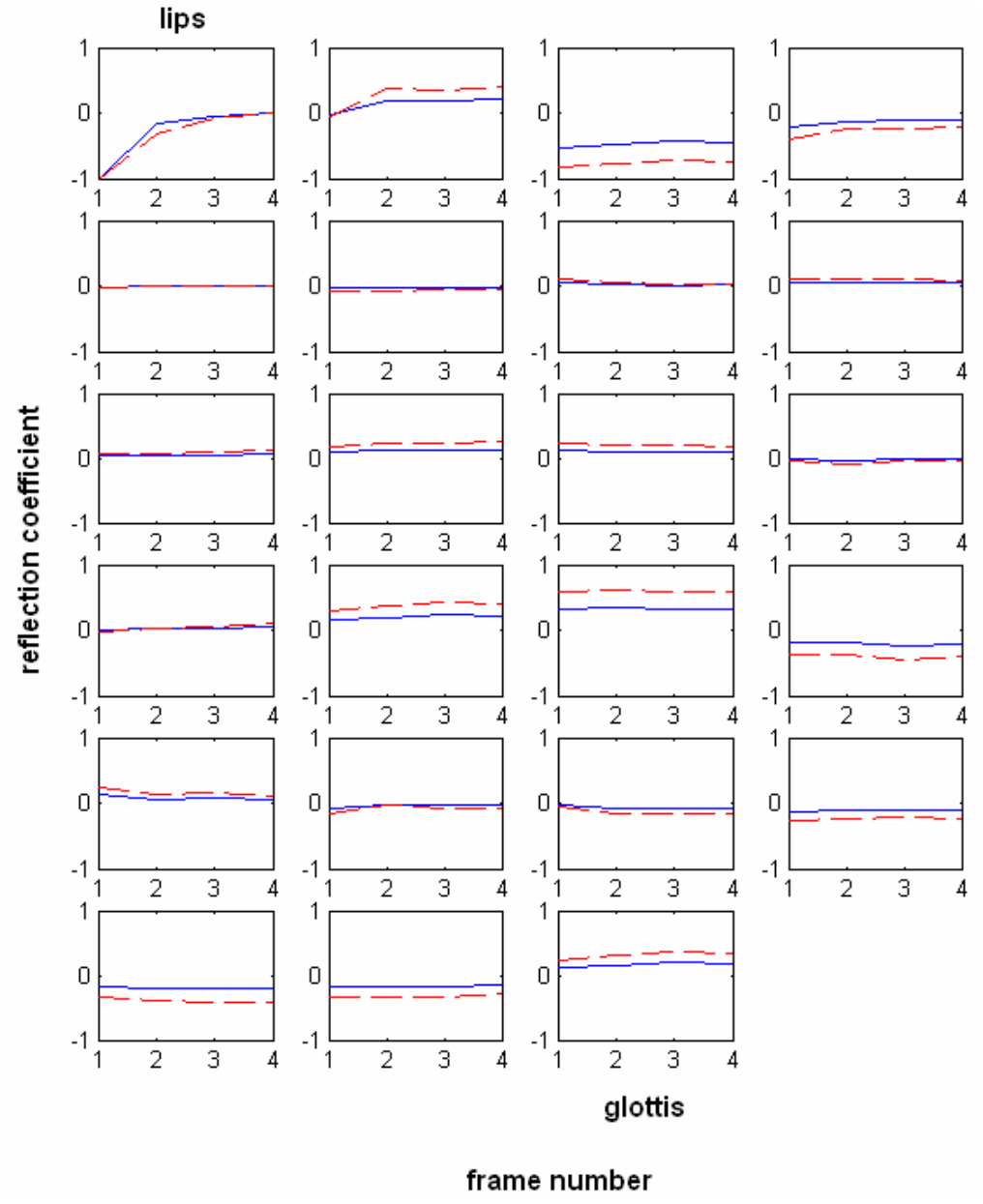

Figure 4-3 : The estimated reflection coefficients of estimated areas for /ba/ utterance with beta equals to 1 (solid lines) and beta equals to 2 (dashed lines).

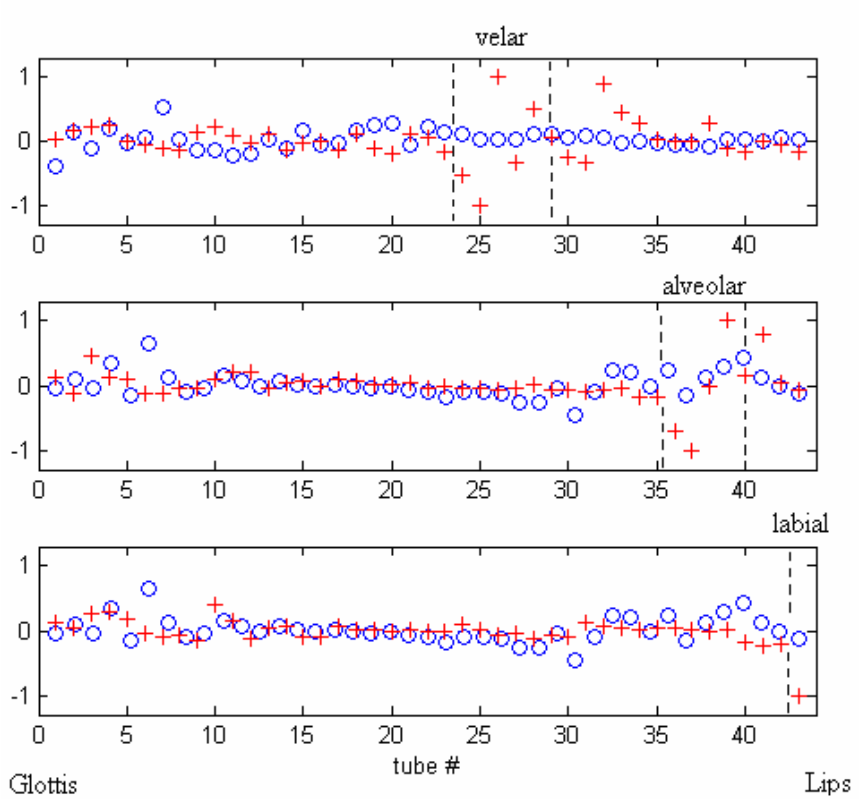

Figure 4-4 : Reflection coefficients of velar (top), alveolar (middle) and labial releases (bottom) (bubbles) with superimposed reflection coefficients of the vowel context (pluses).

In Figure 4-4 shows calculated reflection coefficients of the plosives and the succeeded vowels making use of MRI. The calculation is consistent with the estimation. Remembering Figure 4-1, Figure 4-2 and Figure 4-3, the reflection coefficients were found to be exponential changing functions at constriction regions and except for the constriction regions there were minor changes which can also be modeled by exponential decrease or increase. Examination of Figure 4-4 shows that the reflection coefficients deviate from the following vowel's reflection coefficients greatly only at the constriction regions. The constriction regions corresponding to each transition are shown with dashed lines for better visualization purposes. If we examine the plots one by one, it is observed that the greatest deviation takes place in the first figure which is a velar plosive to back vowel transition (/ga/). In previous chapters the velar transitions were explained to

involve the most major vocal tract changes among all other transitions. To see the difference with another transition, it is enough to examine following plots in the same figure which contains alveolar plosive to front vowel (/di/) and labial plosive to front vowel transitions (/bi/) respectively. In previous chapters, together with velar plosive to back vowel transition, labial plosive to front vowel transition was also mentioned to contain the most simple area changes among all the transitions which are observed from the X-ray investigations. The MRI data is again consistent with this observation because if the remaining two plots are examined, we observe small changes in the reflection coefficients except for the constriction regions that are easily figured out. In addition to the stationary structure, the deviation of these reflection coefficients is observed to be minimum at labial case. The deviation is so small such that even the reflection coefficients of the plosive can be modeled as the same of vowel's with an exception of -1 at the lip end. To demonstrate this rough observation, a simple experiment is performed for /ga/. The plosive reflection coefficients are set equal to the following vowel's reflection coefficients. The reflection coefficients of the plosive, which are at the constrictions, are set to -1 and +1. Then speech is synthesized by changing the reflection coefficient vector exponentially from plosive's coefficients to those of vowels with a suitable time constant. The resonances of vocal tract are plotted for utterance /ga/ in Figure 4-5 (thick lines). For a comparison, typical formants of natural utterances are superimposed with thin lines.

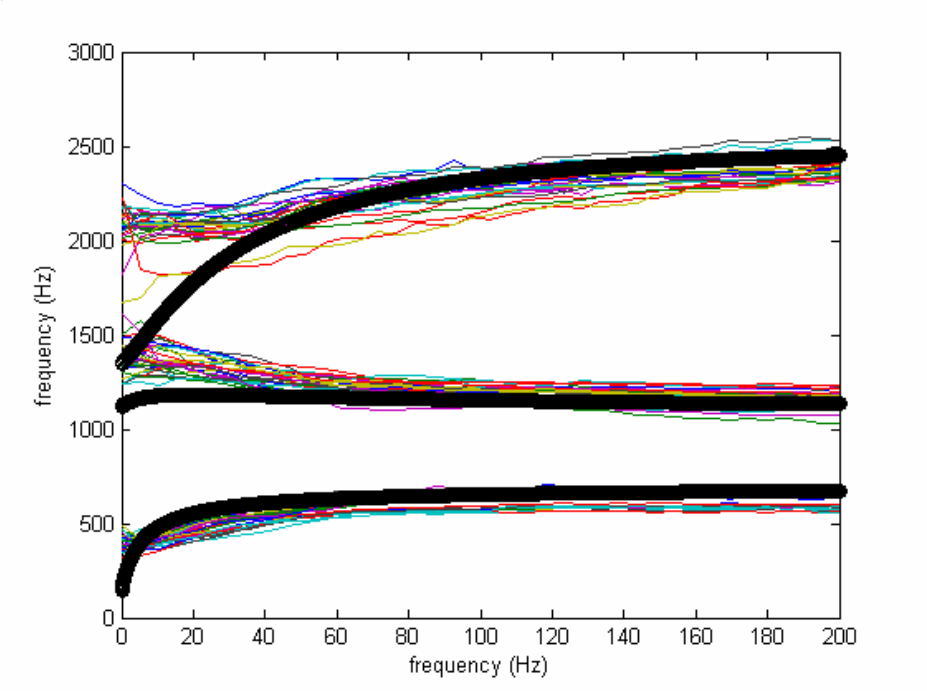

Figure 4-5 : Synthetic vocal tract resonances for utterance /ga/ vs natural ensembles.

Comparing the synthetic formants with natural ones, one observes that, the differences of formants of the natural speech and the synthetic one are not very far away from each other. Indeed, although the rough characters of these transitions are similar, the velar case does not seem to be a good approximation as expected. However the motivation aspect of this experiment is important. So it is decided to follow the second step for further analysis purposes.

In the second step, not only two but all reflection coefficients of the plosives are driven to those of vowels and calculated vocal tract resonance frequencies are plotted in Figure 4-6. Comparing two steps, the similarity in the second is increased because of the number of changed reflection coefficients. However, still differences exist. On the other hand, this is a dummy model which can be a step for verification of observations. Besides the differences are not only the results of the simplicity of the model but also the speaker variation i.e. the vocal tract data and the utterance of /ga/ for which formant ensembles are obtained do not belong to the same speaker. In addition to this, it will be useful to emphasize that the thick

plots are the resonances of the vocal tract without being filtered by excitation, radiation or any other filters whereas the thin ones are the formants of the natural sounds that are calculated by Praat, which also introduces differences to these formant tracks.

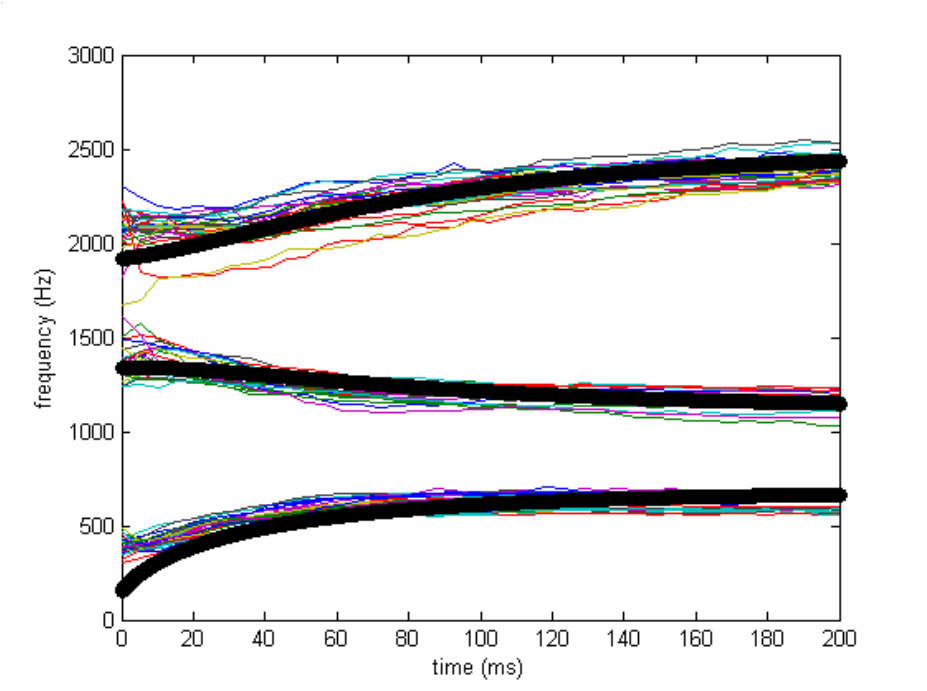

Figure 4-6 : Synthetic vocal tract resonances for utterance of /ga/ vs natural ensembles.

To sum up, the studies up to now seem to be consistent with eac other excluding the minor differences. Thus, a model based upon the areas is proposed with following details.

4.3. Proposed Model for the Plosive to Vowel Transition

- A simple excitation given in Appendix will be applied with a proper length corresponding to the type of the plosive (According to the examinations)
- The amplitude of the excitation may be chosen as a ramp function
- The pitch frequency will be changed exponentially with a proper time constant (According to the examinations).
- For every sample of excitation, the vocal tract area will be updated exponentially to preserve smoothness with a suitable time constant. (Exception exists for velars*).
- A simple lip radiation given in Appendix will be used.

4.3.1. Exception for Velars

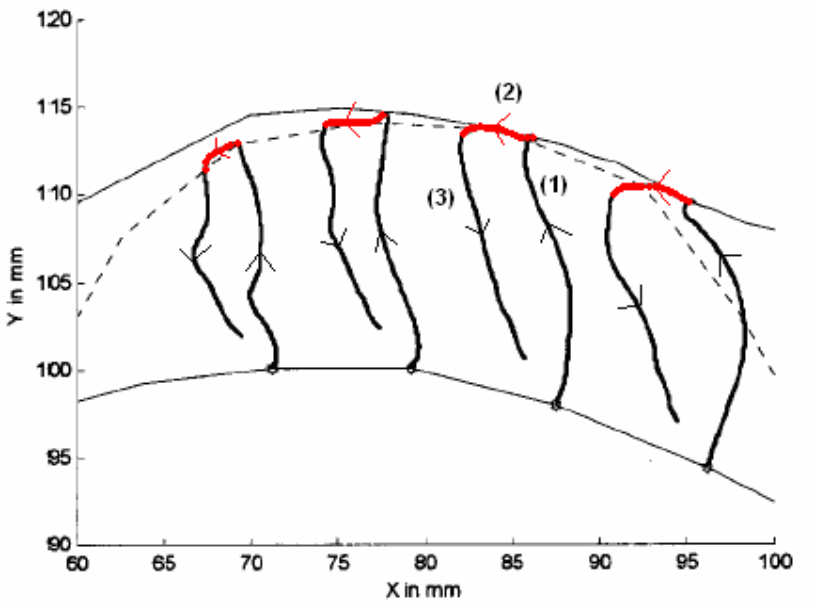

Figure 4-7 : Trajectories of four nodes on the dorsal contour of the tongue in the simulation of /aka/.

In the second chapter, the importance of the VOT region about the perception of the velars, in a back vowel context, is discussed. The importance is related to the convergence of the second and the third formats.

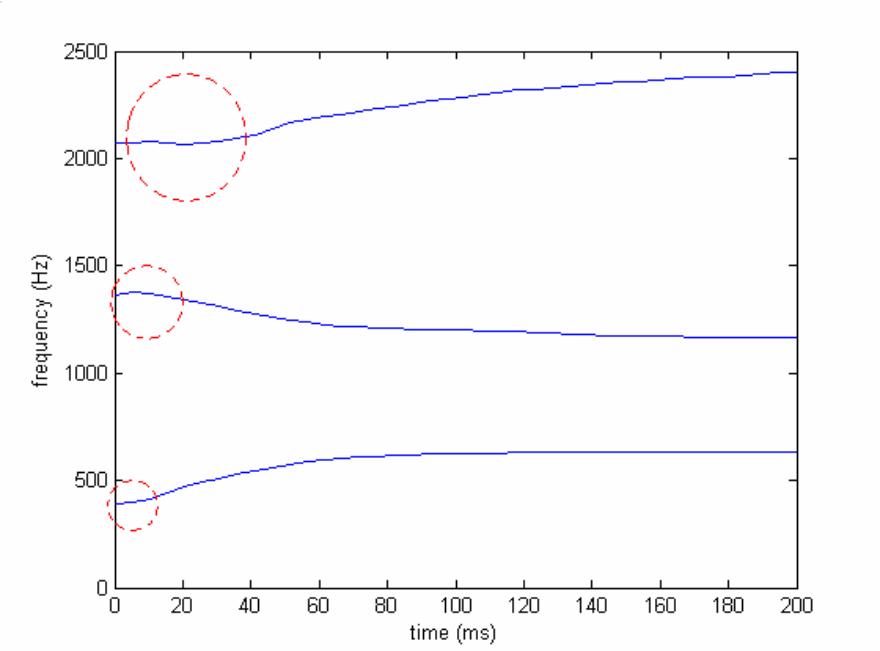

Figure 4-8 : Average formant trajectories for utterance /ga/.

The reason of the large VOT for velar releases can be explained by examining Figure 4-7 and Figure 4-8 carefully. Figure 4-7 shows the trajectories of four nodes on the dorsal contour of the tongue in the simulation of /aka/ by Perrier et al. [12]. The contour, where the tongue rubs the soft palate, is highlighted with red color and marked by (2). Theoretically, VOT is the region which is formed at the interval when the tongue body follows the highlighted region. It is noted that this region is approximately 5-6 millimeters long for each point which is long compared to alveolar and labial cases. In addition to this factor, the body of the tongue is massive compared to other articulators (explained before), therefore VOT is expected to be larger for velars. Another important observation from the figure is that, the sagittal distances do not seem to change marginally during the

VOT interval compared to the regions (2) and (3) as the tongue body moves horizontally. During that interval small openings exist however (For a comparison with a labial case see Figure 4-9. No such trajectory exists especially at snapshots 1, 2, 3, 12, 13 and 14 within the accuracy of 16.6 milliseconds that corresponds to 60 Hz). Voice bar and the glottal excitation radiate through this opening before the release actually forming the VOT region which is rich in the spectral context of the frequency between the second and the third formants of the following vowel (See Figure 3-11 and Figure 4-8). Therefore this interval is important as the spectral context compared to the shorter alveolar and labial cases. Hence the effect of the VOT region should be included to the model.

Figure 4-9 : Sagittal view of /aba/ transition.

4.3.2. Modification of the Model for Velars

As a very simple approach, the horizontal movement of the tongue body can cause very small movements in the vertical i.e. the sagittal direction. Consequently this is the reason of the spectral character of the beginning of the transitions which is spectrally rich at the frequencies between the second and third formants of the

following vowel. Thus, before the rapid release, the constriction can be slightly relaxed more slowly compared to the actual movement at the rest of the transition. A reasonable approximation is to use again an exponential or linear model but double or three times larger than the actual time constant or corresponding linear slope of the rest. After the assumption, a typical area function of the velar constriction with superimposed old area function will be as in Figure 4-10.

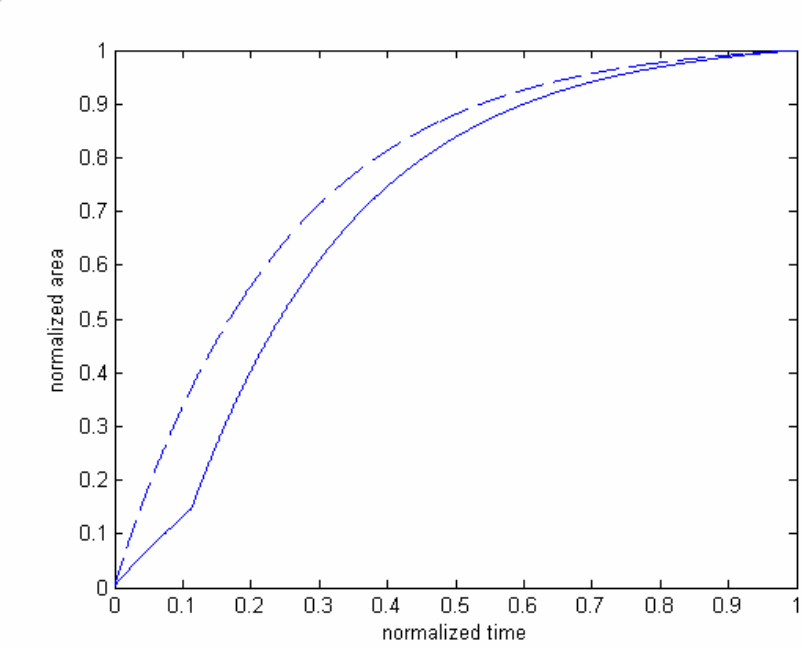

Figure 4-10 : Area function of the velar constriction with VOT taken into account (solid) and not taken into account (dashed).

CHAPTER 5

RESULTS

Three plosive to back vowel and three plosive to front vowel transitions are synthesized according to the proposed model in Chapter 5. Formant structures of the transition, that involve back vowels, are illustrated in the following figures with superimposed natural ensembles.

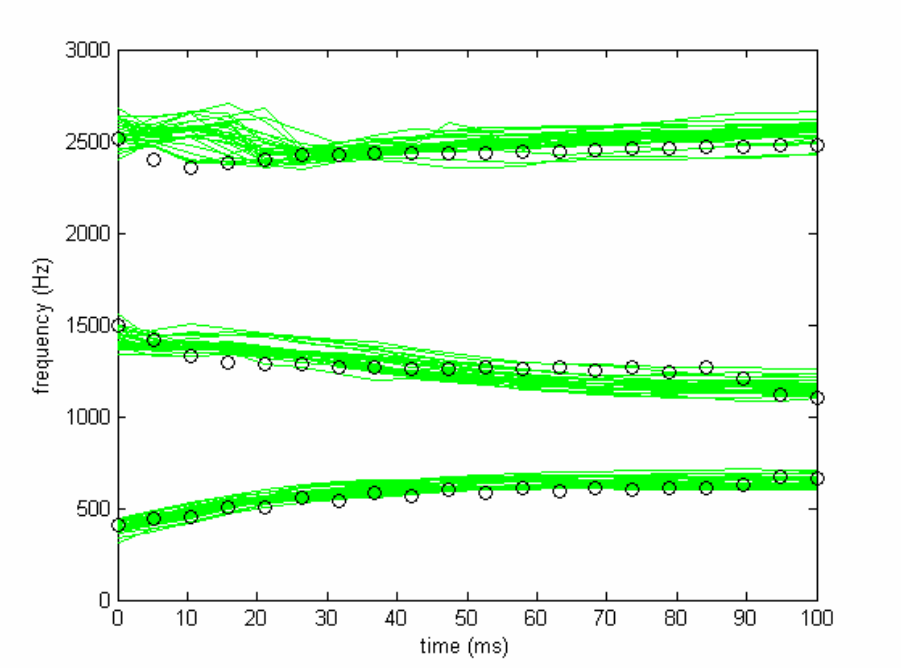

Figure 5-1 : Formant trajectories of synthetic /da/ utterance.

Examining all of the figures it is observed that synthetic and actual formants are in an agreement except for minor differences. These differences can be related to the morphologic differences of the speaker used in X-ray or MRI examinations and the speaker used in the speech ensembles.

In addition to the formant comparison, intelligibility tests of synthesized sounds are done. These six sounds are asked to be recognized by 25 subjects among (b, d and g). For a challenging test, the subjects are asked to recognize three consecutive plosive to back vowel transitions knowing that the sound set is randomly chosen (More clearly subjects are told that the three sounds may be different or all may be the same. However the set always contained three different transitions with back vowels in the first step). In addition, the feedback about the correctness of the decision is not given to subjects during the testing. Same procedure is followed for front vowel case. The results are shown in Table 5-1.

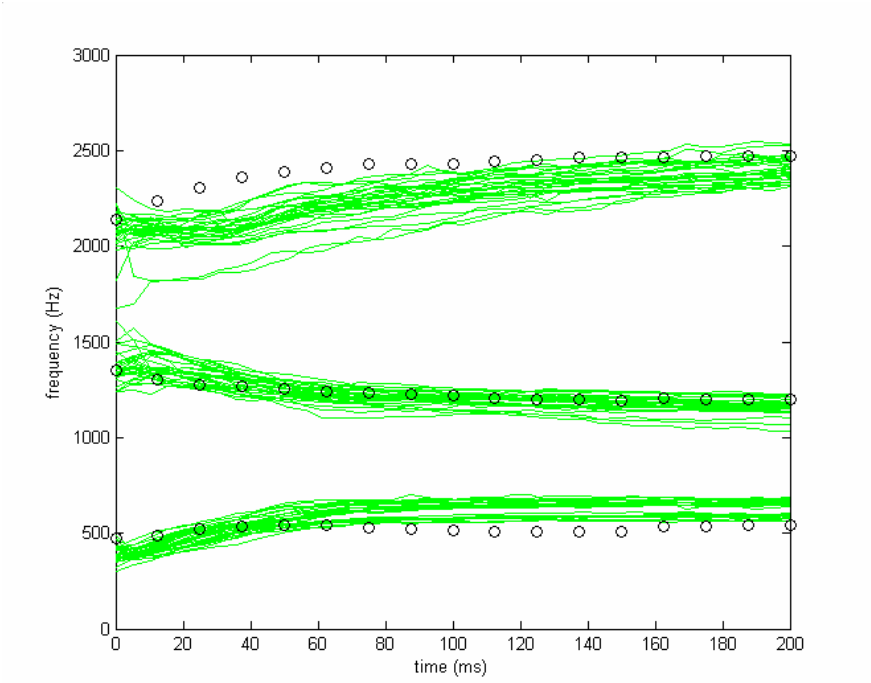

Figure 5-2 : Formants of synthetic /ga/ utterance.

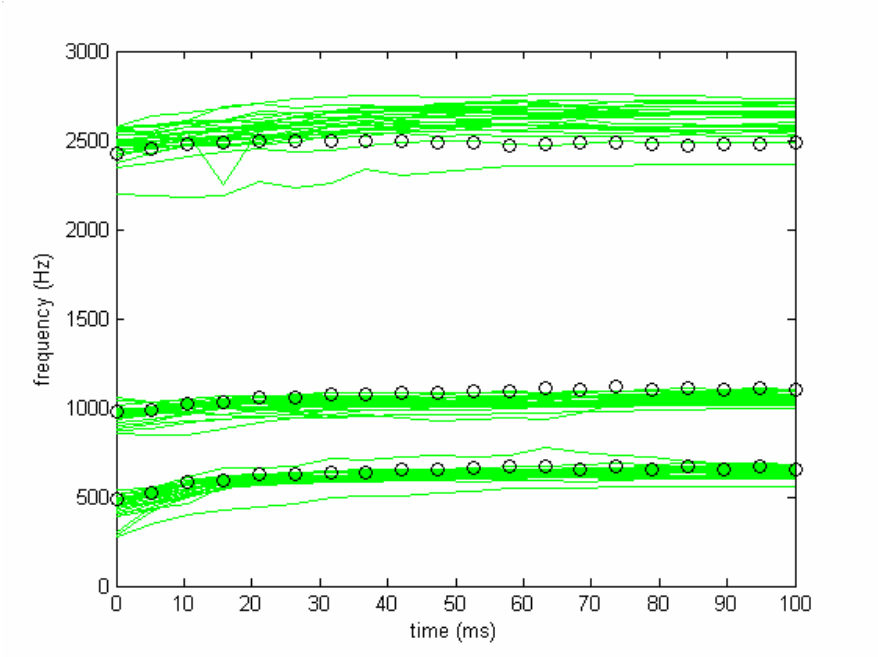

Figure 5-3 : Synthetic /ba/ utterance.

Table 5-1 : Results of the intelligibility test of synthesized transitions.

		Perception					
Actual		**/ba/**	**/da/**	**/ga/**	**/be/**	**/de/**	**/ge/**
	/ba/	**%92**	%8	%0	%0	%0	%0
	/da/	%0	**%96**	%4	%0	%0	%0
	/ga/	%0	%16	**%84**	%0	%0	%0
	/be/	%0	%0	%0	**%100**	%0	%0
	/de/	%0	%0	%0	%0	**%100**	%0
	/ge/	%0	%0	%0	%8	%8	**%84**

Except for a small percentage, the synthesized sounds are almost correctly recognized by the subjects. It is observed that confusion is mostly encountered between velar and alveolar utterances as expected. In previous chapters the confusion is related to the similar formant trajectories. However although such

confusion occurred in the intelligibility tests, the overall recognition rate of the model output is found to be 93 percent which can be assumed to be satisfactory.

In addition, most of the subjects also supplied the feed back that their recognition would not differ even if they are not told to distinguish between these sounds.

As results of the intelligibility tests, it is observed that transitions involving front vowels are better recognized by the subjects. This can be related to the fact that front vowels are more compatible to our model because of the reasons explained in previous chapters. In addition to this observation, velar and alveolar confusions observed in the test are also consistent with the spectral similarities of these sounds which are explained with details in previous chapters.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this study a plosive to vowel transitions are analyzed. In order to achieve this purpose, transitions are examined in different aspects. As a difference from the literature, fast frame rate lips videos are used for labial examinations.

As a result of the examinations, the plosive vowel transitions are observed to be under great dominancy of the following vowel in all aspects. The dominancy is mainly observed in the spectral view which is the result of co articulation in plosive to vowel transitions. In addition to the dominancy of the vowel, exponential behavior of the vocal tract movements are observed. The exponential formant tracks can be related to such a first order dynamics of articulators.

Fricative to vowel transitions are also examined, the plosives are observed to be the inhibited versions of fricatives. The reason behind this observation is related to the various similarities that exist among the plosives and fricatives.

A simple model making use of the vocal tract cross sectional areas is proposed based upon concatenated tube model. Some of the transitions are synthesized according to the proposed model. Intelligibility tests of these synthetic sounds are applied to evaluate the performance of the model to be satisfactory.

However, according to our model we have assumed a single time constant for all vocal tract. This may be a reasonable approach but it will be better to use different time constants for different regions to improve the performance of the model. In addition to the time constant improvement, more accurate sagittal distance to cross section conversions may be done for better results. The inclusion of unvoiced plosive to vowel transitions to that model would be the most satisfactory progress for coverage.

APPENDIX

A.1. Acoustic Theory of Speech Production

(Adapted from Rabiner and Schafer [1])

This part will deal with the mathematical formulations of the process of speech production which are the basis for the analysis and synthesis of speech. First, basic equations of sound propagation will be discussed than the propagation will be applied to a uniform tube model, finally multi tube vocal tract model and its relation to digital filters will be explained. After completing the vocal tract part, excitation source and the radiation parts will be discussed to complete the whole model.

A.1.1. Sound Propagation

Sound is the propagation of acoustic waves which are caused by vibration. The waves travel in air or any other media via the molecules that the media consists of. Thus, the propagation is compatible with the laws of classical physics, especially conservation of mass, conservation of momentum and energy, which are all applicable for sound waves propagating. Using these principles there has been found a set of partial differential equations that describe the motion of the air in the human vocal system. However, the formulation and the solution of these equations are possible under very simple assumptions and simplifications. According to Rabiner and Schafer [1], a detailed acoustic theory should consider the effects of the following factors:

1. Non stationary, time varying vocal tract.
2. Losses due to heat conduction and viscous friction at the vocal tract walls.

3. Softness of the vocal tract walls.
4. Radiation at the lips.
5. Nasal coupling and nasal tract
6. Excitation source at the glottis.

A completely detailed acoustic theory including all above effects is beyond the scope of this thesis, so the most governing factors should be included while omitting the minor effects for a simpler an applicable theory.

The theory of speech production starts with a vocal tract that is modeled as a tube of non-uniform, time-varying, cross-section whose dimensions are very small compared the wavelengths of the audible speech (Speech travels at approximately 330 m/s and frequencies less than 8 kHz have wavelengths 4m and more. Vocal tract length is 17 cm in the average which is approximately %4 of the average wavelength of the speech). Thus, it is reasonable to assume plane wave propagation along the axis of the tube (See Figure A-1). Neglecting heat and viscous losses, discarding the softness of the vocal tract, thus neglecting the radiation of sound while traveling in the vocal tract, also assuming no nasal coupling, a simpler model can be obtained (Discarding the factors 2, 3 and 5 of Rabiner and Schafer [1]). With these simplifications and laws of physics as told before, It is shown that, sound waves in the tube satisfy the following pair of differential equations:

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t} \qquad \textbf{(A.1a)}$$

$$-\frac{\partial u}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p}{\partial t} + \frac{\partial A}{\partial t} \qquad \textbf{(A.1b)}$$

$p = p(x,t)$ is the variation in sounds pressure in the tube w.r.t position x and time t.

$u = u(x,t)$ is the variation in volume velocity flow w.r.t x and time t.

ρ is the density of the air in the tube.

c is the velocity of sound

$A = A(x,t)$ is the "area function" of the tube; i.e., the value of cross-sectional area

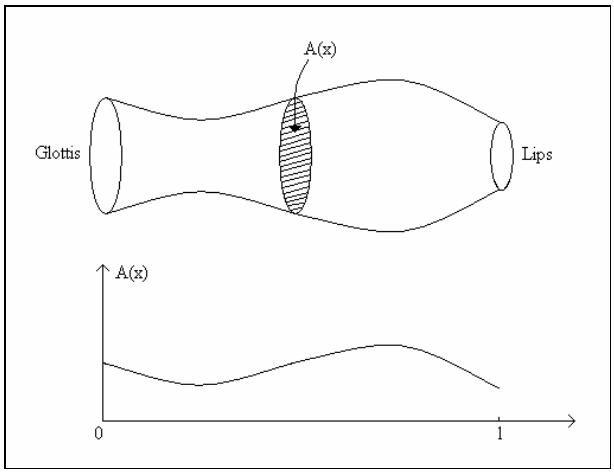

Figure A-1 : Illustration of non-uniform vocal tract and its area function.

These equations are accepted as the basis for speech motion and many useful derivations including speech production, analysis, coding etc. are done using these. However, although the equations are accepted to be the basis for many speech related applications, closed form solutions are not possible except for simplest configurations. However, numerical solutions can be obtained with numerical methods.

Using numerical solution, the pressure and volume velocity can be found for any values of x and t in the region bounded by the glottis and the lips. But, to obtain the solution, boundary conditions must be given at each end of the tube. At the lip end, the boundary condition must account for the effects of sound radiation. At the glottis the boundary condition is imposed by the nature of the excitation.

In addition to the boundary conditions, the vocal tract area function *A(x,t)* must be known for the numerical solution. The area function can be assumed to be constant for stationary sounds, for example vowels, however this is not the case for the non stationary sounds, for example plosives. Therefore exact measurements are needed generally. But detailed measurements of *A(x,t)* are extremely difficult to obtain even for stationary sounds. One approach to such measurements is through use of X-ray motion pictures which will be used in third

chapter. There are many other approaches to find out the vocal tract area functions even when the speaker is speaking in a daily life fashion but such measurements can only be obtained on limited scale and also they can be hazardous to human health (e.g., X-ray) or the method can interfere with the subject's ability to speak (e.g., EPG, electromyography or a velar trace).

To conclude, the area function should be known for numerical solutions. Even if the area function is exactly known, it is hard to solve the differential equations numerically. Therefore, further simplifications should be done for a practical and easier solution.

A.1.2. Uniform Lossless Tube

With a very simple start, it may be useful to consider the vocal tract as a uniform stationary lossless tube of length l (The tube area is constant along x axis and does not change with respect to time t, See Figure A-2). The tube is excited by a back and fort moving piston which generates the air flow i.e. the volume velocity independent of the pressure in the configuration. Although this configuration contains major simplifications and assumptions, it is almost correct for the natural vowel schwa. The reason of the choice of this configuration is the existence of common solution properties with detailed models that will be useful for further derivations. Thus, it is essential to analyze the partial differential equations that are adapted to this configuration.

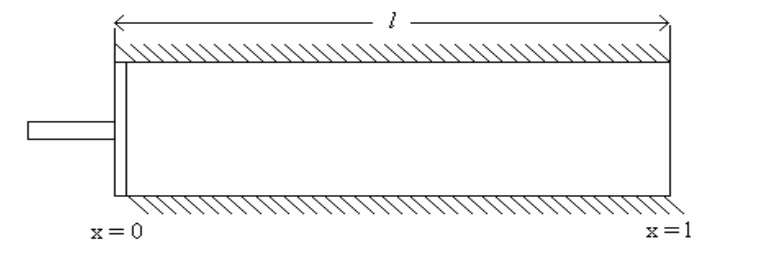

Figure A-2 : Uniform lossless tube.

Remembering partial differential Equations A.1, if *A(x,t)* = *A* is a constant, then the equations reduce to the form:

$$-\frac{\partial p}{\partial x}=\frac{\rho}{A}\frac{\partial u}{\partial t} \quad \textbf{(A.2a)}$$

$$-\frac{\partial u}{\partial x}=\frac{A}{\rho c^2}\frac{\partial p}{\partial t} \quad \textbf{(A.2b)}$$

It can be shown that the solutions to Equations A.2 are in the form:

$$u(x,t)=\left[u^{+}(t-x/c)-u^{-}(t+x/c)\right] \quad \textbf{(A.3a)}$$

$$p(x,t)=\frac{\rho c}{A}\left[u^{+}(t-x/c)+u^{-}(t+x/c)\right] \quad \textbf{(A.3b)}$$

In Equations A.3 the terms $u^{+}(t-x/c)$ and $u^{-}(t+x/c)$ can be interpreted as traveling waves in the positive and negative directions respectively. The relationship between these traveling waves is determined by the boundary conditions.

The frequency-domain representation of this model is obtained by assuming a boundary condition at x = 0 of

$$u(0,t)=u_G(t)=U_G(\Omega)e^{j\Omega t} \quad \textbf{(A.4)}$$

That is, the tube is excited by a complex exponential variation of volume velocity of radian frequency Ω and complex amplitude, $U_G(\Omega)$. Since Equations A.2 are linear, the solution $u^{+}(t-x/c)$ and $u^{-}(t+x/c)$ must be exponentials of the form:

$$u^{+}(t-x/c)=K^{+}e^{j\Omega(t-x/c)} \quad \textbf{(A.5a)}$$

$$u^{-}(t+x/c)=K^{-}e^{j\Omega(t+x/c)} \quad \textbf{(A.5b)}$$

Substituting these equations into Equations A.3 and applying the boundary condition:

$$P(l,t) = 0 \quad \textbf{(A.6)}$$

i.e. assuming zero pressure, at the lip end of the tube and Equation A.4, i.e., the equivalence of volume velocity to the excitation, at the glottis end we can solve for the constants K^+ and K^-. The resulting sinusoidal steady state solutions for *p(x,t)* and *u(x,t)* are

$$p(x,t) = jZ_0 \frac{\sin[\Omega(l-x)/c]}{\cos[\Omega l/c]} U_G(\Omega) e^{j\Omega t} \quad \textbf{(A.7a)}$$

$$u(x,t) = \frac{\cos[\Omega(l-x)/c]}{\cos[\Omega l/c]} U_G(\Omega) e^{j\Omega t} \quad \textbf{(A.7b)}$$

Where

$$Z_0 = \rho c / A \quad \textbf{(A.8)}$$

is by analogy called the *characteristic acoustic impedance* of the tube.

If the solutions are analyzed in a more detailed fashion, it can be figured out that, denominator's zeros correspond to the resonance frequencies of the single tube vocal tract which are also called the *formants*. These frequencies are the peaks in the spectrum that is the energy is mostly concentrated. For numerical values, letting $\cos[\Omega l/c] = 0$ with l = 17.5 cm (the typical length of vocal tract) and c = 35000 cm/s, the resonances are found to be $f_i = 500*(2i-1)$ for $i = 1,2,3....$ Indeed the resonances are compatible to those of the natural vowel schwa whose resonances are approximately 500, 1000 and 2500 Hz respectively (See Figure A-3).

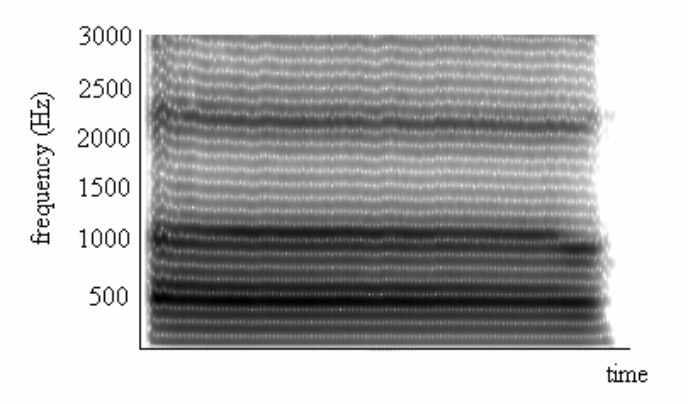

Figure A-3 : The spectrogram of the utterance (natural vowel) /uh/.

To sum up, the starting point of the speech production model, the uniform lossless tube, is a simple but useful example that contains the basis principles for further derivations. From starting, to end of complete model, the equations derived in this section will serve as an origin about sound propagation.

A.1.3. Concatenated Lossless Multi Tube Model

In previous section the simplest model of the speech production is discussed with some details of the important features of the acoustic theory of speech production. To have a complete and more detailed model, sound generation and radiation factors should be integrated in the model with a more sophisticated vocal tract filter. Figure A-4 shows a generic block diagram that is representative of numerous models that have been used as the basis for speech processing so far. These models all have in common that the excitation, vocal tract and radiation features are separated from each other. The radiation and the excitation will be discussed later whereas the vocal tract model will be upgraded.

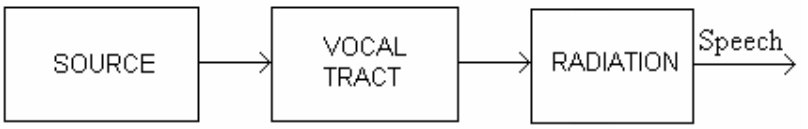

Figure A-4 : Generic Source-Filter Block Diagram.

A commonly used vocal tract model for speech production is based upon the assumption that the vocal tract can be modeled as a concatenation of different sized lossless acoustic tubes, as shown in Figure A-5. The constant cross-sectional areas $\{A_k\}$, along each tube are chosen from the real area function by interpolation method with a proper interpolation scale, If a large number of tubes of short length is interpolated, it can be expected that the resonance frequencies of the concatenated tubes can be close to those of a tube with continuously varying area function. However, since this approximation neglects several losses as told in section 2.1.1, the bandwidths of the resonances may also be expected to be different from those of a more detailed model which includes the losses that the simplified model does not. However, losses can be included to the concatenated tube model at the glottis and at the lips by some extra parameters which may help to represent the resonance frequencies of the speech signals better than the complete lossless case. In this fashion the bandwidths of the resonances will be expected to be closer to the natural sound.

On the other hand, approximating the concatenated tube model from the continuous one, in fact an analog to digital conversion is being done which will be more convenient for implementation and application purposes in today's digitalized world.

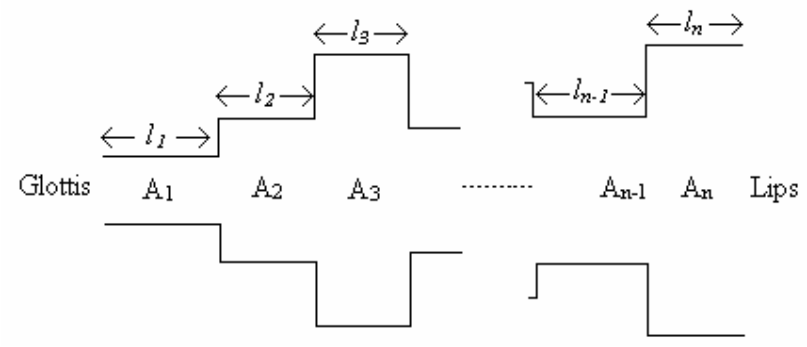

Figure A-5 : Concatenation of n lossless tubes.

A.1.3.1. Wave propagation in concatenated lossless tubes

Since each tube portion is assumed to be lossless, sound propagation in each tube is described by Equations A.2 with appropriate values of the cross-sectional areas. Thus, if the kth tube with cross sectional area A_k is considered, modifying Equations A.3 for the kth part, the pressure and the volume velocity in that tube is in the form

$$p_k(x,t) = \frac{\rho c}{A_k}\left[u_k^{+}(t - x/c) + u_k^{-}(t + x/c)\right] \quad \textbf{(A.9a)}$$

$$u_k(x,t) = \left[u_k^{+}(t - x/c) - u_k^{-}(t + x/c)\right] \quad \textbf{(A.9b)}$$

Where x is the distance measure from the left-hand end of the kth tube $(0 \le x \le l_k)$ and $u_k^{+}(\ )$ and $u_k^{-}(\ )$ are positive and negative traveling waves in the kth portion. To obtain the relationship between the traveling waves in adjacent tubes, the physical principle, the necessity of the continuity of the pressure and the volume velocity in both time and space domain everywhere in the system, can be applied to each portion. The necessity of continuity reveals the importance of using boundary conditions at the each junction. Thus, considering a typical junction between the kth and (k+1)th tubes as shown in Figure A-6, it is enough to apply the continuity conditions for further proceeding, which are:

$$p_k(l_k,t) = p_{k+1}(0,t) \quad \textbf{(A.10a)}$$

$$u_k(l_k,t) = u_{k+1}(0,t) \quad \textbf{(A.10b)}$$

continuity of pressure and volume velocity respectively. Substituting Equations A.9 into A.10 results

$$\frac{A_{k+1}}{A_k}\left[u_k^{+}(t-\tau_k) + u_k^{-}(t+\tau_k)\right] = u_{k+1}^{+} + u_{k+1}^{-} \quad \textbf{(A.11a)}$$

$$u_k^{+}(t-\tau_k) - u_k^{-}(t+\tau_k) = u_{k+1}^{+}(t) - u_{k+1}^{-}(t) \quad \textbf{(A.11b)}$$

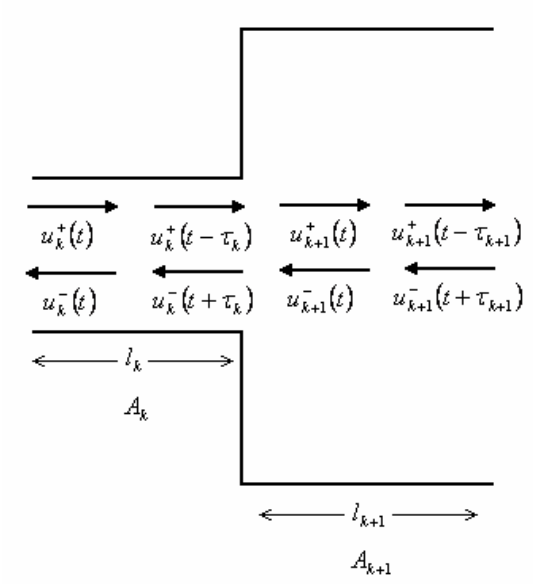

Figure A-6 : Signal flow representation in a typical junction between kth and k+1th tube.

Where $\tau_k = l_k / c$ is the time for a sound wave to travel from the beginning of kth tube to its end. From Figure A-6 it is observed that, some part of the positive going wave is reflected back when it reaches the junction whereas the remaining part keeps propagating to the right. Similarly, a part of negative traveling wave is propagated on to the left while some part is reflected back to the right. Thus,

finding the relation between $u_{k+1}^{+}(t)$ and $u_k^{-}(t+\tau_k)$ in terms of $u_{k+1}^{-}(t)$ and $u_k^{+}(t-\tau_k)$, the propagation character of forward and reverse traveling waves in the kth tube thus in the overall system will be figured out. Solving Equations A.11b for $u_k^{-}(t+\tau_k)$ and substituting the result into Equation A.11a yields

$$u_{k+1}^{+}(t) = \left[\frac{2A_{k+1}}{A_{k+1}+A_k}\right] \cdot u_k^{+}(t-\tau_k) + \left[\frac{A_{k+1}-A_k}{A_{k+1}+A_k}\right] \cdot u_{k+1}^{-}(t) \qquad \textbf{(A.12a)}$$

Subtracting Equation A.11b from Equation A.11a gives

$$u_k^{-}(t+\tau_k) = -\left[\frac{A_{k+1}-A_k}{A_{k+1}+A_k}\right] \cdot u_k^{+}(t-\tau_k) + \left[\frac{2A_k}{A_{k+1}+A_k}\right] \cdot u_{k+1}^{-}(t) \qquad \textbf{(A.12b)}$$

Examining Equation A.12a, it can be figured out that, the quantity

$$r_k = \left[\frac{A_{k+1}-A_k}{A_{k+1}+A_k}\right] \qquad \textbf{(A.13)}$$

is the amount of $u_{k+1}^{-}(t)$ that is reflected at the junction. Thus, the quantity r_k is called the reflection coefficient for the kth junction. It is easily shown that since the areas are all positive, $-1 \le r_k \le 1$. +1 means infinitely increase (total open), whereas -1 means infinitely decrease (total closure).

Modifying Equation A.12 with the definition of r_k, these equations can be reorganized as

$$u_k^{+}(t) = (1+r_k).u_k^{+}(t-\tau_k) + r_k u_{k+1}^{-}(t) \qquad \textbf{(A.14a)}$$

$$u_k^{-}(t+\tau_k) = (-r_k).u_k^{+}(t-\tau_k) + (1-r_k)u_{k+1}^{-}(t) \qquad \textbf{(A.14b)}$$

The equations will be more permanent in minds if they are graphically shown as in Figure A-7. In this figure, general signal flow-graph conventions are used to

represent the multiplications and additions of Equations A.14. Simply, a value on a branch means a multiplication and a node means a summation in which summation of all incoming arrows is equal to summation of all outgoing ones. Clearly, each junction of a system such as that shown in Figure A-5 can be represented by a system such as Figure A-7, Thus for an N tube vocal tract model n-1 junction of the type in Figure A-7 are characterized by n-1 reflection coefficients but since the aim of the vocal tract model is to relate the input to the output, the boundary conditions must be accounted for lips and glottis.

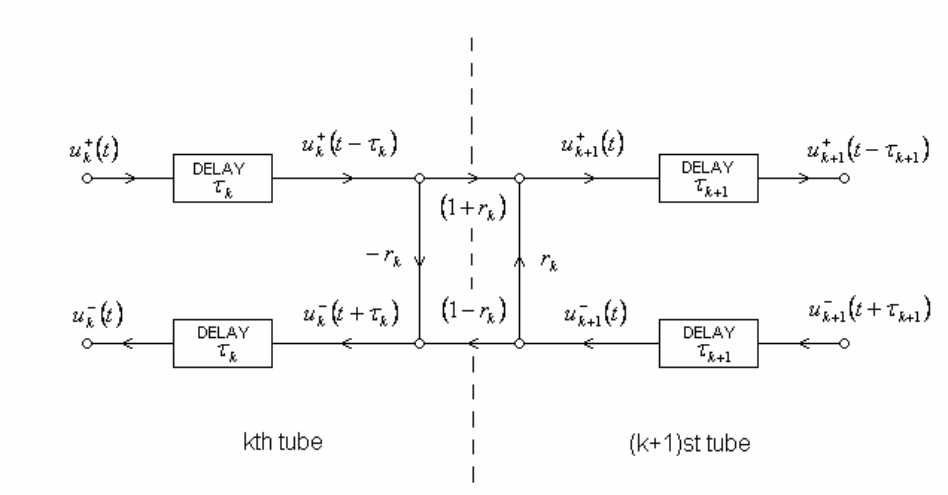

Figure A-7 : Signal flow graph in a typical junction between kth and k+1th tube.

A.1.3.2. Boundary Conditions

Assuming that there are N sections indexed from 1 to N, the starting tube 1 is the one at the glottis. Then the boundary condition at the lips will relate pressure, $p_N(l_N,t)$, and the volume velocity, $u_N(l_N,t)$, at the output of the Nth tube. Assuming the radiation as a complex load, a relation of the form can be obtained.

$$P_N(l_N,\Omega) = Z_L U_N(l_N,\Omega) \qquad \textbf{(A.15)}$$

Assuming Z_L is real, then the time domain relation is obtained as

$$\frac{\rho c}{A_n}\left(u_N{}^{+}(t-\tau_N)+u_N{}^{-}(t+\tau_N)\right)=Z_L\left(u_N{}^{+}(t-\tau_N)-u_N{}^{-}(t+\tau_N)\right) \quad \textbf{(A.16)}$$

Solving for $u_N{}^{-}(t+\tau_N)$ we obtain:

$$u_N{}^{-}(t+\tau_N)=-r_L u_N{}^{+}(t-\tau_N) \quad \textbf{(A.17)}$$

Where the reflection coefficient at the lips is

$$r_l=\left[\frac{\rho c/A_N-Z_L}{\rho c/A_N+Z_L}\right] \quad \textbf{(A.18)}$$

The output volume velocity at the lips is:

$$\begin{aligned} u_N(l_N,t) &= u_N{}^{+}(t-\tau_N)-u_N{}^{-}(t+\tau_N) \\ &= (1+r_l)u_N{}^{+}(t-\tau_N) \end{aligned} \quad \textbf{(A.19)}$$

The effect of this termination as represented by Equations A.17 and A.19 is depicted in Figure A-8. Generally for simplicity and practical aspects, the reflection coefficient is assume to be about 0.71 in speech literature and in this thesis it will be used as this value.

For glottal reflection coefficient, the equation in Rabiner and Schafer [1] will be used.

$$U_1(0,\Omega)=U_G(\Omega)-P_1(0,\Omega)/Z_G \quad \textbf{(A.20)}$$

Assuming again Z_g is real,

$$u_1^+(t) - u_1^-(t) = u_G(t) - \frac{\rho c}{A_1}\left[\frac{u_1^+(t) + u_1^-(t)}{Z_G}\right] \quad \textbf{(A.21)}$$

Solving for $u_1^+(t)$ following relation is obtained:

$$u_1^+(t) = \frac{(1+r_G)}{2}u_G(t) + r_G u_1^-(t) \quad \textbf{(A.22)}$$

Where the glottal reflection coefficient is

$$r_G = \left[\frac{Z_G - \frac{\rho c}{A_1}}{Z_G + \frac{\rho c}{A_1}}\right] \quad \textbf{(A.23)}$$

Equation A.22 can be depicted as in Figure A-9. Again for practical reasons Z_g is found to be very close to 1 by experimental works. In this thesis it is also assumed to be 1.

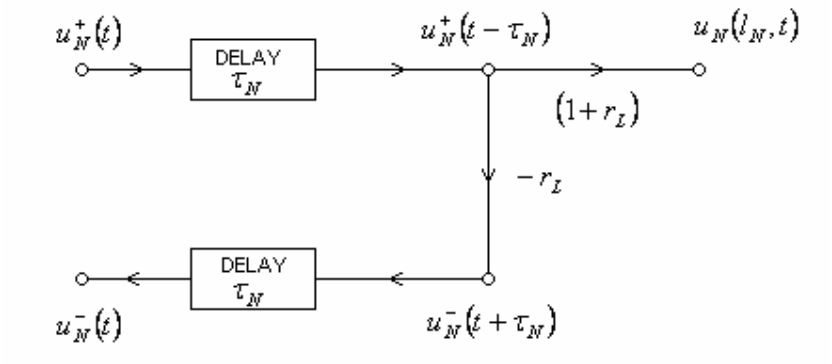

Figure A-8 : Lip end of the concatenated tube model.

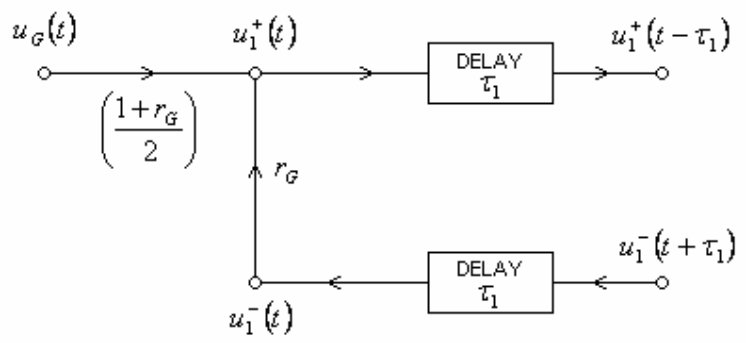

Figure A-9 : Glottal end of the concatenated tube model.

As a simple example, the complete vocal tract diagram with two tubes is shown in Figure A-10. Using Mason's gain formula of the signal flow graphs, the frequency domain response of the system can be shown to be

$$V_a(\Omega)=\frac{U_L(\Omega)}{U_G(\Omega)}=\frac{0.5(1+r_G)(1+r_L)(1+r_1)e^{-j\Omega(\tau_1+\tau_2)}}{1+r_1r_2e^{-j\Omega 2\tau_1}+r_1r_Le^{-j\Omega 2\tau_2}+r_Lr_Ge^{-j\Omega 2(\tau_1+\tau_2)}} \quad \textbf{(A.24)}$$

Several features of Va(Ω) are worth pointing out. First, note the factor $e^{-j\Omega(\tau_1+\tau_2)}$ in the numerator. This represents simply the total propagation delay in the system from glottis to lips. The system function of the system is found by replacing jΩ by s in Equation A.24, with the result

$$V_a(s)=\frac{U_L(s)}{U_G(s)}=\frac{0.5(1+r_G)(1+r_L)(1+r_1)e^{-js(\tau_1+\tau_2)}}{1+r_1r_2e^{-s2\tau_1}+r_1r_Le^{-s2\tau_2}+r_Lr_Ge^{-s2(\tau_1+\tau_2)}} \quad \textbf{(A.25)}$$

The poles of Va(s) are the complex resonance frequencies of the system. There will be an infinite number of resonances because of the exponential dependence upon s. Thus, the formulation should be related to the digital filters for practical aspects as told before.

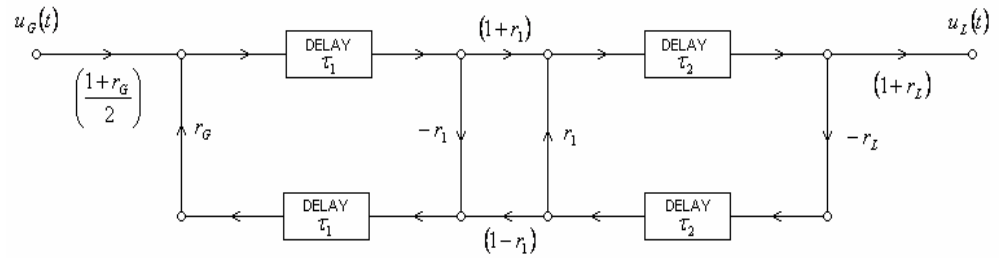

Figure A-10 : Complete flow diagram of two tube model.

A.1.3.2. Relationship to Digital Filters

The form of Va(s) for two tube model suggests that lossless tube models have many properties in common with digital filters. To see this, let us consider a system composed of N lossless tubes $\Delta x = l/N$, where l is the overall length of the vocal tract. Wave propagation in this system can be represented as in Figure A-7 with all delays being equal to $\tau = \Delta x / c$, the time to propagate the length of one tube. It is instructive to begin by considering the response of the system to a unit impulse source, $u_G(t) = \delta(t)$. The impulse propagates down the series of tubes, being partially reflected and partially propagated at the junctions. A detailed consideration of this process will confirm that the impulse response (i.e., the volume velocity at the lips due to an impulse at the glottis) will be of the form

$$V_a(t) = \alpha_0 \delta(t - N\tau) + \sum_{k=1}^{\infty} \alpha_k \delta(t - N\tau - 2k\tau) \qquad \textbf{(A.26)}$$

Clearly, the soonest that an impulse can reach the output is Nτ sec. Then successive impulses due to reflections at the junctions reach the output at multiplies of 2τ seconds later. The quantity 2τ is the time required to propagate both ways in one section. The system function of a such a system will be of the form.

$$V_a(s) = \sum_{k=0}^{\infty} \alpha_k e^{-s(N+2k)\tau}$$

$$= e^{-sN\tau} \sum_{k=0}^{\infty} \alpha_k e^{-s2\pi k} \qquad \textbf{(A.27)}$$

The factor $e^{-sN\tau}$ corresponds to the delay time required to propagate through N sections. The quantity

$$\hat{V}_a(s) = \sum_{k=0}^{\infty} \alpha_k e^{-sk2\tau} \qquad \textbf{(A.28)}$$

is the system function of a linear system whose impulse response is simply $\hat{v}_a(t) = v_a(t + N\tau)$. This part represents the resonance properties of the system. The frequency response $V_a(\Omega)$ is

$$\hat{V}_a(\Omega) = \sum_{k=0}^{\infty} \alpha_k e^{-j\Omega k2\tau} \qquad \textbf{(A.29)}$$

It is easily shown that

$$\hat{V}_a\left(\Omega + \frac{2\pi}{2\tau}\right) = \hat{V}_a(\Omega) \qquad \textbf{(A.30)}$$

This is of course, very reminiscent of the frequency response of a discrete time system. In fact, if the input to the system (i.e., the excitation) is band-limited to frequencies below $\pi/(2\tau)$, then we can sample the input with period $T = 2\tau$ and filter the sampled signal with a digital filter whose impulse response is

$$\begin{aligned} \hat{v}(n) &= \alpha_n \qquad n \geq 0 \\ &= 0 \qquad else \end{aligned} \qquad \textbf{(A.31)}$$

For a sampling period of $T = 2\tau$, the delay of $N\tau$ sec corresponds to a shift of N/2 samples. Note that if N is even, N / 2 is an integer and the delay can be

implemented by shifting the output sequence of the first system, If N is odd, however, an interpolation would be required to obtain samples of the output. This delay would most likely be ignored or avoided in some way (see below) since it is of no consequence in most applications of speech models.

The z-transform of $\hat{v}(n)$ is simply $\hat{V}_a(s)$ with e^{sT} replaced by z. Thus,

$$\hat{V}(z) = \sum_{k=0}^{\infty} \alpha_k z^{-k} \qquad \textbf{(A.32)}$$

a signal flow graph for the equivalent discrete time system can be obtained from the flow graph of the analog system in an analogous way. Specifically, each node variable in the analog system is replaced by the corresponding samples. Also each τ sec delay is replaced by a 1/ 2 sample delay, since $\tau = T/2$.

The 1/2 samples delays in Figure A-11b imply an interpolation half-way between sample values. Such interpolation is impossible to implement exactly, A more desirable configuration can be obtained by observing that the structure of the Figure A-11b. has the form of ladder, with the delay elements only in the upper parts. Signals propagate to the right in the upper part and to the left in the lower part. We can see that the delay around any closed path in Figure A-11b will be preserved if the delays in the lower branches are literally moved up to corresponding branches directly above. The overall delay from input to output will be wrong but this is a minor significance in practice and theoretically can be compensated by the insertion of correct amount of advance (in general $z^{N/2}$). Figure A-11c shows how this is done for the three tube example. The advantage of this form is that the difference equations can be used iteratively to compute samples of output from the samples of the input.

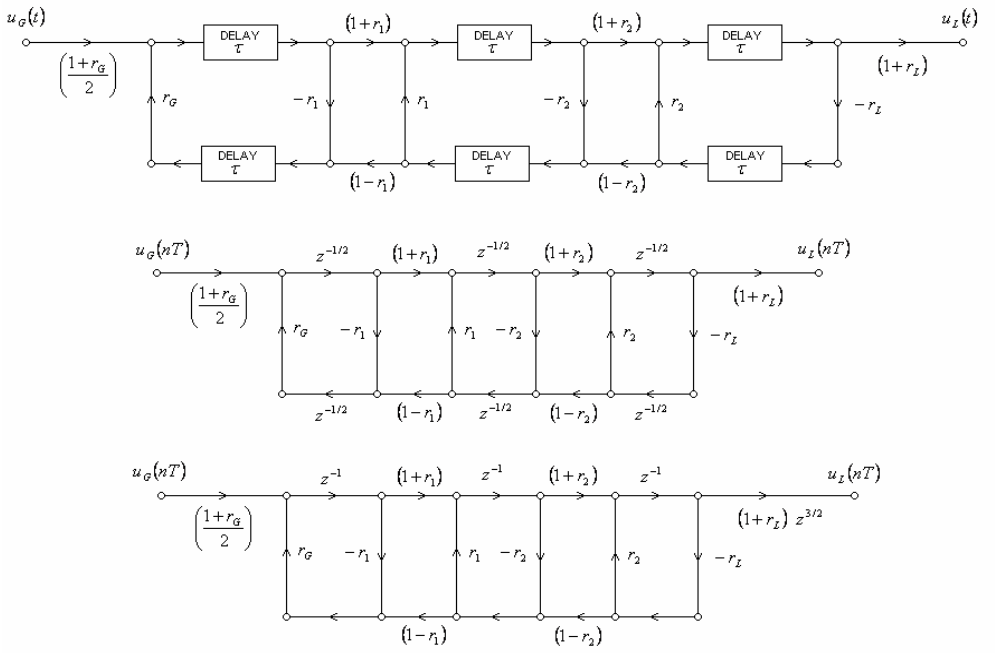

Figure A-11 : a) Signal Flow graph of the concatenated tube model (top) b) equivalent discrete time system (middle) c)simplified discrete time system (bottom).

To demonstrate the vocal tract filter obtained from the analog version, different configurations of vocal tracts are constructed and corresponding reflection coefficients are calculated to form the filter. The resonances for natural vowel schwa, close vowel /i/ and open vowel /a/ are calculated from each configuration. The results are shown in the following figure.

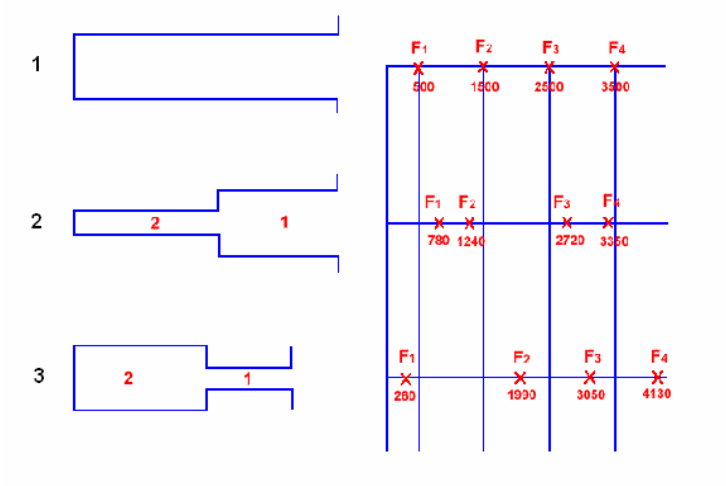

Figure A-12 : The vocal tract configuration and corresponding resonances for /schwa/, /a/ and /i/ respectively.

If the resonances are compared to those of natural sounds, it will be concluded that the multi tube vocal tract model is useful for modeling vowels even with few number of tubes.

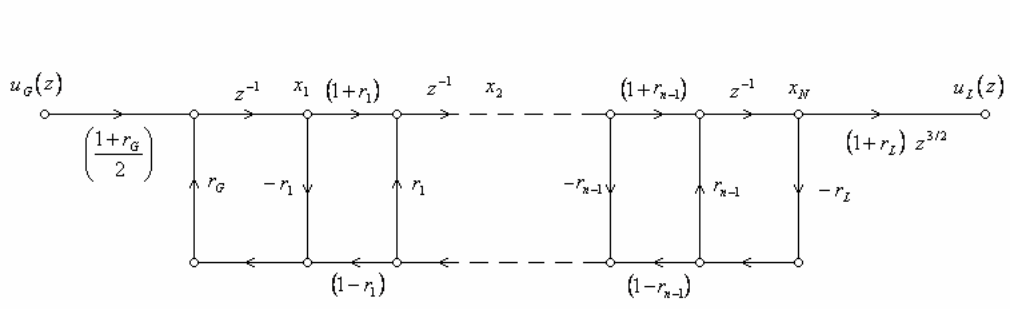

Figure A-13 : Nth order Vocal Tract Filter.

A.1.4. Radiation

So far, the transfer function *V(z)*, which relates the volume velocity at the source to the volume velocity at the lips, is considered. But what the human being

hears is the pressure of the sound which makes the physical impact at the eardrums. So, to have a complete model, the pressure of the sound must be obtained instead of its velocity.

According to experimental works, it is found that, the lips act as a differentiator which converts the volume velocity to pressure by its radiation property. Assuming radiation is independent from the vocal tract model, it can be cascaded to the output of the vocal tract tube. So, the following equation can be written in z domain.

$$P_L(z) = R(z)\, U_L(z) \qquad \textbf{(A.33)}$$

Its character is found to be like high pass filtering operation. In fact, at low frequencies it can be argued that the pressure is approximately the derivative of the volume velocity. Thus, to obtain a discrete-time representation of this relationship, a digitization technique that avoids aliasing should be used. For example, by using bilinear transform method of digital filter design it can be shown that a reasonable approximation to the radiation effects is obtained with:

$$R(z) = R_0\left(1 - z^{-1}\right) \qquad \textbf{(A.34)}$$

i.e., a first backward difference. The crude “differentiation” effect of the first difference is consistent with the approximate differentiation at low frequencies that is commonly assumed.

This radiation “load” can be cascaded with the vocal tract model as in Figure A-14. V(z) can be implemented in any convenient way and required parameters will, of course, be appropriate for the chosen configuration; e.g., area functions for the lossless tube models or formant frequencies and bandwidths for the cascade model.

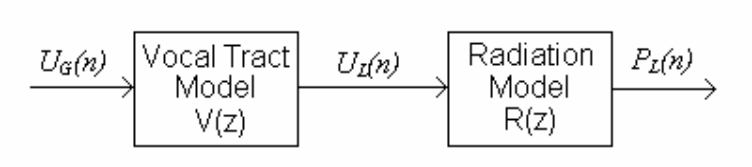

Figure A-14 : Cascaded Radiation Model.

A.1.5. Excitation

To complete the model, a convenient input should excite the system. Recalling that the majority of speech sounds can be classified as either voiced or voiceless, it can be concluded that in general terms what is required is a source that can produce either a quasi-periodic pulse waveform or a random noise waveform.

In the case of voiced speech, the excitation waveform is a raised pulse train A convenient way to represent the generation of glottal wave is shown in Figure A-15a. The impulse train generator produces a sequence of unit impulses which are spaced by the fundamental period. This signal in turn excites a linear system whose impulse response g(n) I has the desired glottal wave shape. A gain control, Av, controls the intensity of the voiced excitation.

The choice of the form of g(n) is probably not critical as long as its Fourier transform has the right properties. Rosenberg, in a study of the effect of glottal pulse shape on speech quality, found that the natural glottal pulse could be replaced by a synthetic pulse waveform of the form:

$$\begin{aligned} g(n) &= \frac{1}{2}\left[1-\cos(\pi n / N_1)\right] && 0 \le n \le N_1 \\ &= \cos\left(\pi(n-N_1)/2N_2\right) && N_1 \le n \le N_1 + N_2 \\ &= 0 && otherwise \end{aligned} \qquad \textbf{(A.35)}$$

This wave shape is very similar in appearance to the pulses. Figure A-15a shows the pulse waveform and its Fourier transform (Figure A-15b) magnitude for typical values of N_1 and N_2 (in Figure A-15 they are chosen to be 40 and 10

respectively). It can be seen that, as would be expected, the effect of the glottal pulse in the frequency domain is to introduce a low-pass filtering effect.

Since g(n) in Equation A.35 has finite length, its z-transform, G(z) has only zeros. An all-pole model is often more desirable. Good success has also been achieved using a two pole model for G(z).

For voiceless sounds the excitation model is much simpler. All that is required is a source of random noise and gain parameter to control the intensity of the unvoiced excitation. For discrete models, a random number generator provides a source of flat spectrum noise. The probability distribution of the noise samples does not seem to be critical.

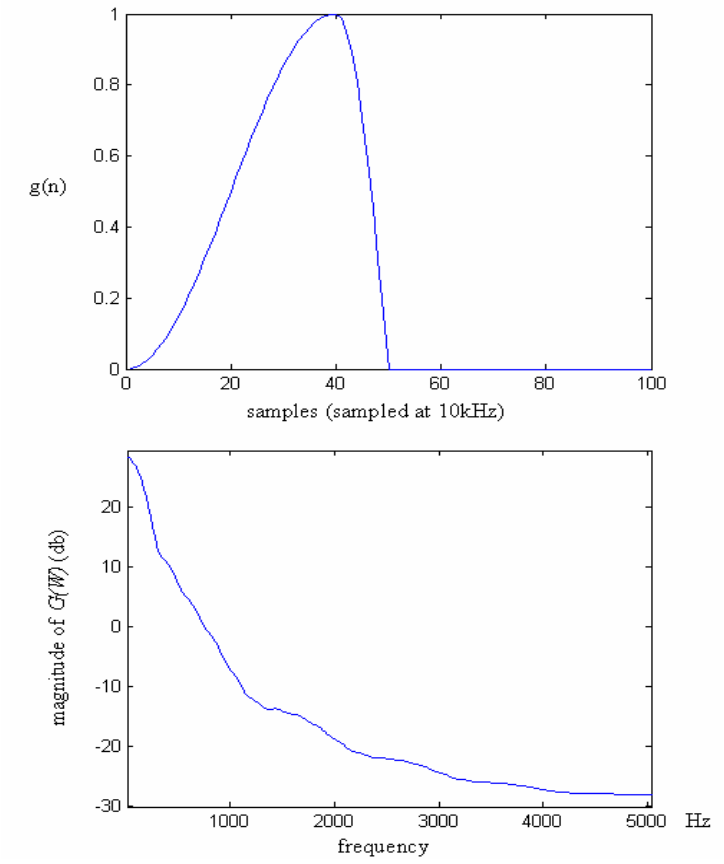

Figure A-15 : a) Time waveform of the glottal pulse b) its spectrum.

A.1.6. Complete Model

The concatenated tube model is obtained with its source and radiation. The cascade connected model can be shown as in Figure A-16. The voiced / unvoiced switch separates the voiced and unvoiced excitation source whereas the radiation and vocal tract model are common for both types of sound. Also an impulse train generator is needed for voiced sounds whereas random noise generator is essential for unvoiced ones. In addition to the pulse train generator for voiced sounds, also a glottal pulse shaper is needed to filter the impulse train which is not essential for unvoiced case. For both cases, the intensity of the sound is controlled by two different amplifiers to have the desired power at the output.

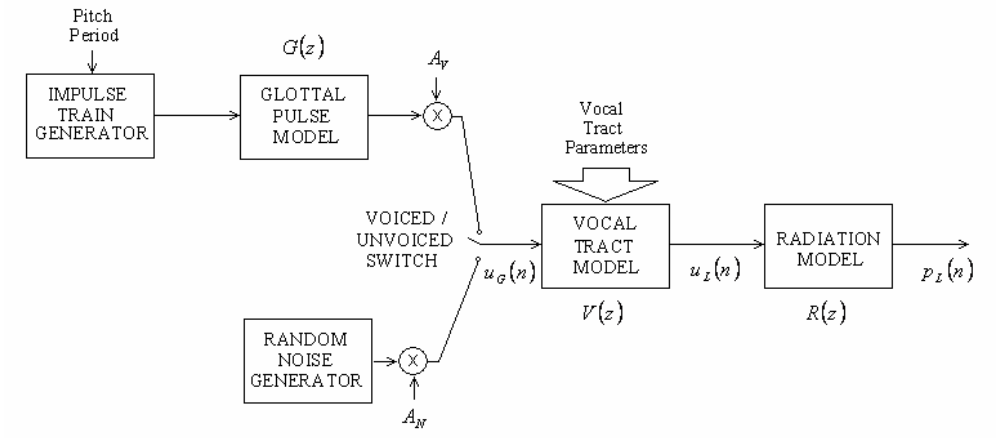

Figure A-16 : The complete concatenated tube block diagram.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, “Digital Processing of Speech Signals”, Prentice Hall, Englewood Cliffs. NJ., 1978.

[2] J. L. Kelly, Jr. and C. Lochbaum, “Speech Synthesis”, Proc. Stockholm Speech Communications Seminar, R.I.T., Stockholm, Sweden, 1962.

[3] M. R. Portnoff, “A Quasi-One-Dimensional Digital Simulation for the Time-Varying Vocal Tract”, M. S. Thesis, Dept. of Elec. Engr., MIT, Cambridge, Mass., 1973.

[4] P. J. B. Jackson, “Characterisation of plosive, fricative and aspiration components in speech production”, Ph.D. thesis, Dept. Electronics & Comp. Sci., Univ. of Southampton, UK, 2000.

[5] P. J. B. Jackson, “Acoustic cues of voiced and voiceless plosives for determining place of articulation”, Proc. of the Workshop on Consistent and Reliable Acoustic Cues for sound analysis, CRAC 2001, 19-22, Aalborg, Denmark, 2001.

[6] Mohammed, M.A.S., “Dynamic measurements of speech articulators using magnetic resonance”, Ph.D. thesis, Dept. Electronics & Comp. Sci., Univ. of Southampton, UK, 1999.

[7] Mermelstein, P., “Articulatory model for the study of speech production”, Journal of the Acoustical Society of America, Vol. 53, pp. 1070-1082, 1973.

[8] Öhman S and Stevens K N, "Cineradiographic studies of speech: procedures and objectives", Journal of the Acoustical Society of America, Vol 35, pp. 1889, 1963.

[9] Zue, V. W., “Acoustic characteristics of stop consonants: A controlled study”, D. Sc. thesis, MIT, 1976.

[10] Suchato, A., “Classification of stop consonant place of articulation”, Ph.D. Thesis, MIT, 2004.

[11] P. Perrier, L. J. Boë, and R. Sock, “Vocal tract area function estimation from mid-sagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients”, J. Speech Hearing Res., Vol. 35, pp. 53-67, 1992.

[12] Perrier P., Payan Y., Zandipour M. and Perkell J, “Influences of tongue biomechanics speech movements during the production of velar stop consonants: A modeling study”, Journal of the Acoustical Society of America, Vol. 114, pp. 1582-1599, 2003

[13] Soquet A. ;Lecuit V; Metens T. ;Demolin D, “From Sagittal Cut to Area Function: An RMI Investigation”, Proc. ICSLP, pp. 1205-1208, 1996.

[14] Mrifric. Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst and L. J. Gerstman, “Some Experiments on the Perception of Synthetic Speech Sounds”, Journal of the Acoustical Society of America, Vol. 24, pp. 597-606, 1952.

[15] Greenwood, A.R., Goodyear, C.C., Martin, P.A, “Measurements of vocal tract shapes using magnetic resonance imaging”, IEEE Proc., Vol139, pp. 553-560, 1992.

[16] Baer, T, Gore, J. C., Gracco, L. C., and Nye, P. W, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels", Journal of the Acoustical Society of America, Vol. 90, pp. 799-828, 1991.

[17] Shrikant S. Narayan, Aber A. Alwaan, and Haker K., "An articulatory study of fricative consonants using magnetic resonance imaging", Journal of the Acoustical Society of America, Vol. 98, pp. 1325-1347, 1995.

[18] Praat version 4.3.29, Copyright (1992-2005), Paul Boersma and David Weenink.

[19] http://www.ling.mq.edu.au

[20] Stevens K. N., "Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations", Journal of the Acoustical Society of America, Vol. 50, pp. 1180-1192, 1971.

[21] Stevens K. N, Klatt D. H., "The Role of Formant Transitions in the Voiced-Voiceless Distinction for Stops", Journal of the Acoustical Society of America, Vol. 55, pp. 653-659, 1974.

[22] Stevens K. N, "Models for production and acoustics of stop consonants", Speech Communication, Vol. 13, pp. 367-375, 1993.

[23] M. Halle, G. W. Hughes and J. P. A. Radley, "Acoustic properties of stop consonants", Journal of the Acoustical Society of America, Vol. 29, pp. 107-116, 1957.

[24] Fischer and Jorgensen E., "Acoustic analysis of stop consonants", Miscellanea Phonetica, Vol. 2, pp. 42-59, 1954.

[25] Story B. H., Titze I. R., and Hoffman E. A., "Vocal Tract Area Functions from Magnetic Resonance", Journal of the Acoustical Society of America, Vol. 100, pp. 537-554, 1996.

[26] Maeda. S., "On the conversion of vocal tract X-ray data into formant frequencies", Murray Hill, NJ: Bell Laboratories, 1972.

[27] Fant, G., “Acoustic theory of speech production”, S-Gravenhage; Mounton, 1960.

[28] J. L. Flanagan, “Speech Analysis, Synthesis and Perception”, 2nd Ed., Springer-Verlag, New York, 1972.

[29] Heinz J. M. & Stevens K. N., “On the relations between lateral cineradiographs, area functions and acoustic spectra of speech”, Proceedings of the Fifth International Congress of Acoustic, A44. Liege, 1965.

[30] Sundberg, J., “On the problem of obtaining area function from lateral X-ray pictures of the vocal tract. Speech Transmission Laboratory-Quarterly Progress and Status Report”, Vol. 1, pp. 43-45, University of Stockholm, 1969.

[31] Fant, G., “Formants and cavities”, Proceedings of the Fifth International Congress of Phonetic Sciences, pp. 120-141, MUnster, Basel: Karger, 1964.

[32] Virtual Dub – Copyright (1998 – 2003) by Avery lee.

[33] L. L.. Beranek, “Acoustics”, McGraw-Hill Book Co., New York, 1954.

[34] Alwaan A., “Modeling Speech Perception in Noise: The Stop Consonants as a Case Study”, RLE Technical Report, MIT, 1992.