VOICE CHARACTERISTICS
Parameterization of Articulatory Movements Based on Cascaded First-order Systems and its Application

K. Ogata and Y. Sonoda

Faculty of Engineering, Kumamoto University, Kumamoto, 860-8555 Japan

The purpose of this paper is to show several examples of the application of cascaded first-order systems to describing articulatory behavior. The first example is analysis of articulatory movements. The model was applied to analyzing the articulatory characteristics between two speakers. A comparison of the model parameters between the speakers showed different strategies in articulation to the increase of speaking rate. The second example is reproduction of articulatory behavior by combining several articulatory trajectories based on the model. Since the proposed model can extract information on articulatory timing from observed data, the trajectories of four measuring points on the tongue surface, which were separately collected, were combined based on information on the timing. The reproduced movements provided useful information on the change in the shape of the tongue surface during articulation.

CASCADED FIRST-ORDER SYSTEMS AND ARTICULATORY MODELING

The authors have studied articulatory modeling based on cascaded first-order systems[1]. The modeling method can describe the trajectories of articulatory movements with a relatively small number of parameters. In this paper, we show several examples of its application to analysis and reproduction of articulatory behavior.

In our modeling method, the cascaded first-order systems[2] are used as a simplified model of higher order complex systems. The impulse response of the model is given by

\[ h_a(t) = \frac{A}{(n-1)!} T^n e^{-\frac{t}{T}}, \quad t > 0 \]  

(1)

where \( n \) is the number of cascaded components, that is, the order of the model, \( T \) is the time constant, and \( A \) is the amplitude. Figure 1 schematically shows articulatory movement patterns based on the step responses of the cascade; displacement, velocity, and acceleration. As shown in Fig.1(b), a step input as a hypothetical motor command to the cascade produces an ascending motion. Successive ascending and descending motions are obtained by combining two responses as shown in Fig.1(c). Based on this idea, observed articulatory movements are parameterized through a fitting procedure, in which optimum values of the model parameters are obtained by a least square method.

EXPERIMENTS

The first example of the application is analysis of articulatory behavior. Tongue, jaw and lower lip movement data within the midsagittal plane were collected by means of the magnetic and optical sensing system developed by the authors[3].

Figure 2 shows an example of lower lip and jaw movements for one speaker during the utterance /eppep/. For the lower lip, the movements of the lower lip relative to the jaw were evaluated. Based on the modeling method in Figure 1, the observed movement pattern is closely approximated by the sequence of the responses from four hypothetical commands of which input instants are indicated by four vertical lines in the figure. The calculated movement pattern almost overlaps with observed one.

In order to analyze articulatory characteristics between two speakers, behavior of the model parameter \( A \), which is related to movement amplitude, was evaluated when speaking rate changed. Figure 3 shows a comparison of the effect of speaking rate between the two speakers. In the figure, the values defined as

\[ R = \frac{\sum_{i=1}^{2} \frac{A_{F_{\text{fast}}}}{A_{F_{\text{normal}}}} - 1}{2} \]

(a) Cascaded first-order systems.

(b) Response pattern. (c) A combination of two response pattern

FIGURE 1. Articulatory behavior based on cascaded first-order systems.
are shown for the utterances /epepe/ and /etete/. The value is defined as the ratio of two ratios for primary articulator \(A_{P,Fast}/A_{P,Normal}\) and secondary articulator jaw \(A_{J,Fast}/A_{J,Normal}\). The primary articulator corresponds to the lower lip for /epepe/ and tongue for /etete/, respectively. Each ratio indicates the ratio of the value \(A\) at the fast speaking rate to that of the normal rate. This value shows effects of speaking rate on the displacement of the primary articulator and jaw. Therefore, in the case of extreme reduction of amplitude at jaw compared to the primary articulator, the value \(R\) of Eq.(2) becomes positive. In the opposite case, the value of it becomes negative. In the figure, speaker KO shows positive values in both utterances and articulates in such a manner as the jaw presented a stronger reduction of its movement amplitude than the tongue and lower lip to the increase of speaking rate. In contrast, speaker YS shows negative values in the utterance /epepe/ and shows that the change in speaking rate affected the lower lip movement compared to the jaw. In the case of /etete/, YS shows small positive and negative values. Thus, evaluation of the parameters reveals different strategies in articulation between the speakers to the increase of speaking rate.

The second example is reproduction of articulatory behavior by combining parameterized articulatory trajectories. This method is based on the advantage that the proposed model can extract information on articulatory timing from observed data as shown in Figure 2. Figure 4 shows an example of the reproduction of the articulatory movement for /a/ to /i/ transition in the utterance /aia/. The contours of the tongue surface are shown every 10 ms by connecting four points by a spline function. The trajectories of the four points of 10, 20, 30, and 45 mm from the tongue tip, which were separately collected and parameterized by Eq.(1), were combined based on the information on the extracted timing between each tongue point and the jaw. As shown in the figure, the reproduced movement patterns provide useful information on the change in the shape of the tongue surface during articulation.

ACKNOWLEDGMENTS

Part of this work was supported by Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology.

REFERENCES


FIGURE 2. Observed and calculated movement patterns for the lower lip and jaw during the utterance /epepe/ at the normal speaking rate for speaker KO.

FIGURE 3. Comparison of the reduction of movement amplitude defined as Eq.(2) between two speakers.

FIGURE 4. Reproduced tongue movement patterns for /a/ to /i/ transition in the utterance /aia/.
Studying Articulatory Effects Through Hypercube Sampling of the Articulatory Space

S. Ouni and Y. Laprie

This paper presents an original study of compensatory properties of the vocal tract. Indeed, due to the high dimension of the articulatory space, exploring all the compensatory effects of the vocal tract is a hard task, even if we could use real data (MRI, x-ray). For these reasons, we use a codebook representing the sampling of the articulatory space. Sampling such a space is difficult because of the non-linear nature of the mapping between articulatory and acoustic spaces. We linearize the mapping by decomposing the articulatory space into regions, the mapping is quasi-linear in. An exploration method of the codebook is developed, in order to study articulatory variability. We present different vocal tract configurations of some vowels and their articulatory characteristics.

INTRODUCTION

This paper presents an original study of compensatory properties of the vocal tract. It is known that normal speech production exploits compensatory effects [2]. In fact, the vocal tract configuration necessary for the production of a vowel can be realized by many different combinations of individual articulatory positions. Thus, the vocal tract can reach an invariant acoustic target by coordinating articulators to compensate for each other. Studying the articulatory variability is generally done by observing real data (MRI, x-ray) as it is done in [4], for instance. While using real data is the best way to study articulatory variability, there are many backwards related to the technique: x-ray is harmful to health, the MRI technique constrains the speaker to adopt abnormal articulator positions while speaking which may alter the interpretation of the experiments, and of course, the quantity of these real data is always limited, because of the high dimension of the articulatory space.

For these reasons, we present a new method to study the articulatory variability. As in [1], an articulatory model is used to carry out the study. We use an articulatory synthesizer which implements Maeda’s [3] articulatory model and describes the vocal tract by seven parameters. This synthesizer is used to generate a codebook which represents the mapping between articulatory and acoustic domains. This codebook is a sampling of the articulatory space. Sampling such a space is difficult because of the non-linear nature of the mapping between articulatory and acoustic spaces. In fact, in some articulatory regions a small perturbation of the articulatory parameters could give rise to large acoustic modifications. Therefore, if the sampling of an articulatory region is not adapted to the acoustic behaviour of the articulatory model, very important acoustic effects could be hidden. Our codebook construction method overcomes this problem and enables the study of all the possible compensatory phenomena of the vocal tract.

HYPERCUBIC SAMPLING

The main idea is to discretize densely only in the regions where the mapping is highly non-linear. Each linear region is represented by a seven dimensional hypercube. First, we assume the articulatory space is contained in a hypercube. If the mapping is not sufficiently linear in a hypercube, this hypercube is decomposed into sub-hypercubes. Thus, the overall structure is a hierarchy of hypercubes. The refinement procedure is repeated until the mapping between the articulatory space and the acoustic space can be considered linear in a hypercube (details of the algorithm construction can be found in [5]).

INVERSION PROBLEM

As the dimension of the articulatory space is seven (the seven parameters of the articulatory model) and the dimension of the acoustic space is three (the first three formants, as we are studying vowels), retrieving the articulatory parameters from acoustics is not trivial. We have developed a method exploiting this codebook by recovering the possible articulatory vectors for each acoustic entry of the signal to be inverted. Details of the method can be found in [6]. We just present the principle of the method. For each acoustic entry, all the hypercubes whose acoustic image contains the acoustic entry is considered. We exploit the gradient expression at a particular point \( P_0 \), in the hypercube (the center for instance):

\[ F - F_0 = \nabla F (P - P_0) \]  

where \( F \) is the acoustic entry, \( VF \) is the gradient of \( F \) calculated at \( P_0 \) and \( F_0 \) is the acoustic vector corresponding to \( P_0 \). \( P \) is the articulatory vector we are looking for. Resolving this matrix system is performed by the SVD (singular value decomposition) method which returns a particular solution and an orthonormal base of the null space associated to this system. As a result, we have a complete specification of the solution set. To retrieve all
the solutions we have to sample the null space. The precision of the solutions depends on the quality of the sampling. Experiments showed that our method gives good results [6]. As we will see in the next paragraph, this method provides all the possible articulatory configurations corresponding to a particular vowel. So an exhaustive study of the compensation can be carried out.

**ARTICULATORY VARIABILITY**

The codebook and exploration method enable the determination of all the possible vocal tract configurations for an acoustic entry. Here, we deal with french vowels. We use our method to recover all the articulatory configurations of the vocal tract. For these vowels, we present the case of the vowel [u]. We obtained about 12000 vocal tract configurations which can be organized in three classes. In Fig. 1, (a), (b) and (c), representatives of the three classes are shown. In this classification, we have considered the constriction place, as it is very influential for the acoustic characteristics [7]. We distinguish the velar [u] (a), the palatal [u] (b) and the pharyngeal [u] (c), depending whether the place of the constriction is the velum, the palate or the pharynx. We note the same constriction place can occur for several jaw positions (a) and (d). These two configurations present the same place of constriction, with two different jaw positions. Lip and, to a lesser extent, larynx compensate the acoustic effects of the jaw variability. In Fig. 2, we present another study for the vowel [y]. Among all the articulatory configurations found, we can extract two classes, which are slightly different in terms of the place of constriction. The first class presents the constriction at the back of the hard palate and the second class in the front of the hard palate. The first class is accompanied by a large lip aperture and the second by lip closure. These two examples clearly show how vocal tract exploits the articulatory compensatory properties to reach the same acoustic target. In these examples, the compensation intervenes between jaw position and lip aperture.

**REFERENCES**

Effects of Glottal Asymmetry on Modes of Phonation

Fariborz Alipour\textsuperscript{1} and David Berry\textsuperscript{2}

\textsuperscript{1}Department of Speech Pathology and Audiology, University of Iowa, Iowa City, IA 52242, USA
\textsuperscript{2}UCLA School of Medicine, Division of Head and Neck Surgery, Los Angeles, CA 90095, USA

ABSTRACT

By applying tension asymmetries in a finite-element model of vocal fold vibration, a variety of modes of phonation were simulated. The biophysical model incorporated tissue mechanics, glottal aerodynamics, and acoustics of the vocal tract. For simplicity and speed, glottal aerodynamics were modeled with a modified Bernoulli equation. Vocal tract acoustics were modeled using a wave reflection and transmission-line approach. An increase in tension in one fold generally increased the oscillation frequency of both folds, which can be explained by the mechanism of eigenmode entrainment, a common phenomenon in nonlinear oscillators. However, the vocal fold with the higher tension had a higher mucosal wave velocity and a smaller amplitude of vibration.

INTRODUCTION

Glottal asymmetry is a common irregularity that has been observed in patients with many types of voice disorders involving asymmetries between the left and right vocal folds. One obvious example is unilateral vocal fold paralysis. Because of motion differences between the two folds, glottal asymmetry may influence the glottal airflow, pressure distributions, and the glottal resistance [1]. The changes in the stiffness between the two folds can be observed by the difference in the mucosal traveling wave velocity [2]. One major result of asymmetry appears to be the increase of phonatory instabilities leading to bifurcations and chaos [3,4].

While videostroboscopic measurements in patients with vocal fold paralysis can help identify the problem and its severity and provide estimates of mucosal wave velocity [5], direct measures of tissue stiffness and their influence on human phonation cannot be studied using this method. One possible approach might be the use of a canine excised larynx model, where asymmetric tension can be induced and phonatory and aerodynamic effects can be studied [6]. Another approach is the use of a computational model that mimics the human larynx. Such a computational model is used in this study. It is based on a finite element solution of viscoelastic waves in a continuum, and has been developed, refined, and validated in our laboratory over the last 20 years [7,8].

RESULTS

A variety of outputs from the simulation helped depict its oscillatory behavior. Nodal coordinates were written into sequential files (frames) to make movies and visualize the oscillations. Glottal waveforms including subglottal pressure, glottal area, volume velocity, glottal amplitude and output pressure described the phonatory characteristics of the model. However, examination of the glottal waveforms did not appear to capture helpful information regarding asymmetries. The most useful output for comparison seemed to be the coordinate waveforms of selected nodes. Figure 1 shows the oscillations of a node on the medial layer near the edge of the vocal fold (node number 3), when the right fold had twice as much tension as the left fold. The solid lines represent the position coordinates of node 3 on the right vocal fold, and the dashed lines the corresponding coordinates on the left fold. As shown,
the right fold had smaller horizontal (X) amplitudes than the left fold. This is consistent with asymmetric oscillations observed in excised larynx experiments [6].

![Figure 1](image1.png)

**Figure 1.** Coordinate waveforms of node number 3 in the medial coronal layer. The lower traces show horizontal vibrations and the upper traces show vertical vibrations. Solid lines indicate oscillations of the right fold and dashed lines indicate oscillations of the left fold.

Increasing the tension of one fold increased the oscillation frequency of both folds. As shown in Figure 2, a Fast Fourier Transform (FFT) of the left and right coordinate waveforms revealed that the two folds oscillated at the same frequency. However, the right fold with more tension had a smaller amplitude of vibration in the horizontal direction (Figure 1, solid line).

The phase plot of this oscillation represented a limit cycle. However, when the tension ratio exceeded 2.5, harmonics appeared in the phase plot. Furthermore, for high pressures (20 cm H₂O) and high tension ratios, some frequencies appeared that were not harmonics of the fundamental. With the presence of two or more independent frequencies, the phase plot now resembled a torus. The acoustic output sounded rough and oscillations became unstable. Thus, bifurcations occurred in the vocal fold oscillations, i.e., jumps to qualitatively different oscillation patterns occurred for a small change in tension ratio. Such nonlinear phenomena are commonly observed in vocal fold vibration. They occur because of nonlinearities associated with vocal fold collision, tissue viscoelasticity, and pressure-flow relations in the glottis. Such phenomena have been observed previously in human phonation and excised larynx experiments [3,4,6].

![Figure 2](image2.png)

**Figure 2.** Frequency spectrum of the right and left horizontal displacement waveforms of the node 3 as shown in Figure 1.

**ACKNOWLEDGMENTS**

This work was supported by grant number R01 DC03566 from the United States National Institutes of Health.

**REFERENCES**

Interactions between Emotional Content of Pictures and Acoustic Features of Speech

R. Ruiz a, R. Da Silva Neves b, C. Martinot c, S. Vautier b

a LA.R.A., b C.E.R.P.P., c L.D.C.C.: Université de Toulouse-Le Mirail, 5 allées Antonio Machado, 31058 Toulouse cedex 1, France (E-mail: robert.ruiz@univ-tlse2.fr).

Vocal manifestations of picture-induced emotion exist in everyday life. The nonlinguistic information of the speaker’s emotional state is investigated here in a laboratory experiment where the correlation between a few acoustic measures on the speech signal and the emotional rating of pictures is studied. The experiment and the first results are presented.

INTRODUCTION

Part of the manifestations of emotion is conveyed by speech. Previous studies have shown that stress and/or emotion modify the speech signal under laboratory and real conditions [1]. Simulation of emotion has been done to study the perception of emotional stress [2, 3]. Various situations have been used to study vocal emotion (psychomotor tasks, cockpit voice recordings, stress tests, etc …). Here, it is induced by pictures. Each one of them is characterized by an “emotional score” composed of an arousal and a pleasure rating between 1 and 9 [4].

A coherent or incoherent text to pronounce is superimposed on the images. And prior to the test subjects have to fill in psychological questionnaires to determine their anxiety state and trait. Therefore multiple correlations are possible not only between acoustical features but also with the emotional scores, the coherence of the sentences and the anxiety.

Time measures performed on the speech signal are studied first. Indeed, analysis of the literature seems to indicate that some of them are the most correlated with various emotionally situations. And unlike the spectral ones they are not phoneme-dependant. They can constitute a reliable indicator basis for the study of emotion detection.

EXPERIMENT AND RESULTS

Pictures belong to the International Affective Picture System (I.A.P.S) database [4]. They are marked by their coordinates in an arousal versus pleasure plane. In four areas of the plane, 24 images are selected (6 by region) : the maximum arousal / maximum pleasure one (A+/P+), the minimum arousal / minimum pleasure one (A-/P-), the A+/P- and the A-/P+ ones. Pictures are proposed in a random order.

Each picture is associated with both a French consistent sentence and an inconsistent one. Each sentence starts by a stop Consonant - Vowel - stop Consonant pseudo-word (C1.V.C2) corresponding to an element of the picture. For example if there is a river on the screen, the coherent sentence (C.S) is: “tip is a river” (in French) and the incoherent one (I.S) is “tip is a lake” (in French). C.S and I.S are randomly distributed but in equal proportion. C.V.C structures are composed of stop consonants [ p ], [ k ], [ t ]. The choice of stop-consonants is based on the fact that their start is more easily time-detectable than the major part of the other ones because of their impulsive nature. They are associated with vowels [ a ], [ i ], [ y ] in the following combinations : [tip], [pit], [pak], [kap], [tyk], [kyt]. These three vowels are chosen to cover a large frequency domain for future spectral analysis.

A sound signal (like a “bip”) is synchronized with image appearance to give the time reference. Indeed, speakers have to pronounce the sentence after they heard the signal. The picture appears during 6 seconds, then the sentence appears during 4 seconds and the “bip” is emitted 6 seconds after the sentence have disappeared from the computer screen.

Before trial pictures, subjects had to utter a list of the six C1.V.C2 structures used (Phasis 1). In the final phase (Phasis 3) of the experiment, they also have to repeat all the sentences, showed again on the screen, but without the images. These two sets of phonetic material form the pre- and post-state of rest”. During the Phasis 2 the 24 pictures are seen and their associated sentences uttered.

D.A.T recordings are done in a sound studio with an A.K.G cardioid prepolarized condenser microphone (model C420) with a behind-the-neck headband for hands free use. A foam windscreen is used and the microphone is placed near the corner of the mouth to avoid pop noise.

Time measures on the signal are performed by Matlab programs. They are:

1/ the inverse of the mean fundamental period of the low-pass filtered vowel V signal (i.e the mean fundamental frequency) noted $1/T_0$ (in Hz);
2/ the standard deviation of the mean fundamental period computation noted SD_T0 (in seconds) ;  
3/ the jitter of the fundamental period of the low-pass filtered vowel signal (in %).

The first results are reported for only 6 speakers (all students, 3 men and 3 women) and four sets of two pictures both belonging to one of the four groups of the arousal / pleasure plane (Table 1). Table 2 shows the significance of the variations between Phasis 1 (state of rest) and Phasis 2 for all the combinations of the three acoustic features and the four picture groups. For the others comparisons the probabilities of H_0 are of the same magnitude.

### Table 1. Results of vocal signal measures are averaged for the 6 speakers studied (standard deviations are into brackets). In the Phasis 2 and 3 results depend on the group the two pictures belong to (A+/P+, A+/P-, A-/P+, A-/P-).

<table>
<thead>
<tr>
<th>Phasis</th>
<th>T0 (Hz)</th>
<th>SD_T0 (s)</th>
<th>Jitter (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/T0</td>
<td>203.16</td>
<td>0.33</td>
<td>4.69</td>
</tr>
<tr>
<td>A+/P+</td>
<td>205.38</td>
<td>0.15</td>
<td>2.74</td>
</tr>
<tr>
<td>A+/P-</td>
<td>200.52</td>
<td>0.21</td>
<td>4.61</td>
</tr>
<tr>
<td>A-/P+</td>
<td>200.31</td>
<td>0.30</td>
<td>5.20</td>
</tr>
<tr>
<td>A-/P-</td>
<td>204.92</td>
<td>0.19</td>
<td>3.26</td>
</tr>
</tbody>
</table>

### Table 2. Probabilities p of H_0 for the t-test between Phasis 1 and Phasis 2.

<table>
<thead>
<tr>
<th></th>
<th>1/T0 (%)</th>
<th>SD_T0 (%)</th>
<th>Jitter (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+/P+</td>
<td>0.323</td>
<td>0.322</td>
<td>0.325</td>
</tr>
<tr>
<td>A+/P-</td>
<td>0.358</td>
<td>0.359</td>
<td>0.499</td>
</tr>
<tr>
<td>A-/P+</td>
<td>0.494</td>
<td>0.453</td>
<td>0.363</td>
</tr>
<tr>
<td>A-/P-</td>
<td>0.472</td>
<td>0.492</td>
<td>0.472</td>
</tr>
</tbody>
</table>

**DISCUSSION**

Examination of the results leads to the following temporary conclusions.

Numerical variations of the acoustic features mean values between the state of rest (Phasis 1) and the emotional state (Phasis 2) have not of a large extent and are not significant.

Standard deviation of fundamental periods of the segmented vowel and jitter seems to decrease when emotional content is maximum (A+/P+ set of pictures).

Even if the increase of both numbers of speakers and pictures can lead to a better significance of the variations, the characteristics studied will not probably become reliable for the experiment because of the great overlap between the variability inter-speaker and the variations observed. Perhaps the experimental procedure has a masking effect because an important part of the emotion can have disappeared before the “bip” gives the order to speak. A complementary study is needed to ensure that the choice of this reaction time is not too large.

Examination of individual results is more encouraging because largest variations are observed on the speech signal of subjects who seem to less control their emotion before they utter.

**CONCLUSION**

The experiment needs to be developed with the entire set of images and the twenty five subjects remaining. More investigations about time cues (reaction time, phoneme duration etc . . . ) and spectral cues are also needed with multiple correlation analysis involving the coherence of the sentence and the state of anxiety of the speaker. Results cannot suggest for the moment that vision can acts upon phonation like audition do it. Since the acoustic properties of an utterance are linked to the acoustic properties of a heard sound (Lombard effect for example), they are also probably linked to the visual properties and the emotional content of a viewed picture.

**REFERENCES**

Voiced / Unvoiced Classification Using Subband Crosscorrelation Analysis

Z. Lachiri\textsuperscript{a}, N. Ellouze\textsuperscript{b}

Laboratoire LSTS, Département de Génie Electrique, Ecole Nationale d’Ingénieurs de Tunis
Campus Universitaire, BP 37, 1002, Le Belvédère, Tunis, Tunise
\textsuperscript{a}zied.lachiri@enit.rnu.tn \textsuperscript{b}N.Ellouze@enit.rnu.tn

In this paper we propose a new voiced / unvoiced classification algorithm in noisy environment using an appropriate speech subband decomposition. Speech sound is considered to be a signal which contains both high/low frequency components and short/long duration sounds. Considering its mathematical property and the capability to model speech sounds, the wavelet packet is well suited to this type of expansion. The classification is achieved by generating a correlation model of different subbands signals derived from a tree structured filter banks. To investigate the accuracy of the proposed technique, we conduct experiments using the TIMIT speech database. We add to these speech signals real world noise at various SNR. Experimental results show the accuracy of the proposed technique especially in low SNR’s ($\leq 10\,\text{db}$).

WAVELET PACKET SUBBAND DECOMPOSITION

Wavelet transform \cite{2} was recently introduced as an alternative technique for analysing non stationary signal. The continuous wavelet transform of signal $s$ relative to the basic wavelet is given by:

$$W_\psi s(a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} s(t) \psi^* \left( \frac{t-b}{a} \right) \, dt \quad (1)$$

where $a, b \ (a, b \in R; a \neq 0)$ are respectively the translation and scale parameters. Discrete time implementation of wavelet is based on a tree structure which uses a single basic building block repeatedly until the desired decomposition is accomplished. This basic unit uses techniques of multi-rate signal processing and consists of a low and a high pass filter followed by a down-sampling unit. Unlike the wavelet transform, which only partitions the frequency axis finely toward the low frequency, the wavelet packet transform divides not only the low, but also the high frequency subband, resulting in tree structured filter bank which we call a wavelet packet filter bank \cite{4}. This transformation creates a division of the frequency domain that represents the signal optimally with respect to the applied metric while allowing perfect reconstruction of the original signal.

SPEECH SUBBAND DECOMPOSITION

Speech can generally be classified as voiced, unvoiced, or mixed. The voiced sound is frequency limited signal which has most of the energy in the low frequency range, less than 1 KHz, whereas the energy of unvoiced speech is usually concentrated at the high end of frequency scale ($\geq 3$KHz) \cite{3}. It is known that most of the speech signal power is contained around the first formant frequency which doesn’t exceed 1 KHz and doesn’t below 100Hz approximately. In addition, pitch frequency lies in normal speech between 80 and 500 Hz. Based on these behaviours, we suggest to decompose the speech signal into 8 subband wavelet packet tree which is constructed by cascading the basic two channel filterbank into various levels. Each of these subband signals contains only restricted frequency information due to inherent bandpass filtering. The filter bank that implements the wavelet packet decomposition is given in (Figure 1).

FIGURE 1. 8 subband wavelet packet tree covering 0 – 8KHz and their parameters: Center frequency (Hz) and Bandpass (KHz).

VOICED UNVOICED CLASSIFICATION

The speech signal is highly correlated in case of speech and is not correlated in case of silence. This fact make it possible to track the uncorrelated portions and extract the pure speech segments. In effect, any transition
between a silence and voiced sound or unvoiced sound can be identified by the Subband Crosscorrelation Analysis (SUB-CRA) between different subband signals obtained via wavelet packet subband decomposition.

The algorithm begins by splitting the speech signal into smaller windows. Each window is passed through an appropriate filter bank to extract the wavelet packets parameters. The SUB-CRA is performed using different filters responses (Fig 1). Filter 1, filter 2 and filter 3 are selected to detect the voiced segments and filter 6, 7 and 8 are selected to locate the unvoiced segments. To generate the crosscorrelation function, a simple interpolation technique is used to insert points between the wavelet packets parameters to expand them in each frequency band to the window length. The frames of all crosscorrelation parameters are concatenated, then the absolute value of the points is taken and smoothed to extract the envelope function. Consequently, we obtain 5 crosscorrelation function $CR_{1-2}$, $CR_{2-3}$, $CR_{1-3}$, $CR_{6-7}$ and $CR_{7-8}$ and we choose two crosscorrelation function, which have the maximum of energy among $CR_{1-2}$, $CR_{2-3}$, $CR_{1-3}$ and $CR_{6-7}$, $CR_{7-8}$.

The discrimination of the speech signals (voiced and unvoiced) from noise is conducted using a comparison with an appropriate threshold, which is generated exploiting the first frames of the appropriate crosscorrelation function. To obtain the logical series, the crosscorrelation function is compared with the noise threshold. The element in the series can take a value of 0 or 1 if the threshold of the noise is respectively greater or inferior than the correlation sample.

**EVALUATIONS**

Speech signals used in the experiments were uttered by male and female speakers and obtained from TIMIT corpus. The test set consisted of a total of 465 frames of data. The experiments were conducted by adding real world noise (White, factory, car and F16) with different SNR. The other details in the experiments are as follows: window size 512 point and frame shift 256 point. The results depicted in Figure 2 show good detection performance.

**REFERENCES**

Observation of Sound Pressure Waveform and Vibration Acceleration Waveform on the Nose for Voiced Consonants

Takayoshi Nakai, Shoji Yokoyama, Ryuki Nagasaka and Keizo Ishida

Department of Electric and Electronic Engineering, Faculty of Engineering, Shizuoka University
3-5-1 Johoku, Hamamatsu, 432-8561, Japan    e-mail:tdtnaka@ipc.shizuoka.ac.jp

This paper describes individuality for utterance of voiced consonant. We measured sound pressure waveform and vibration acceleration waveform on the nose, VA. For buzz of initial voiced consonant, VA has large amplitude and individuality of amplitude of VA is large. It is shown that buzz sound mainly radiates from the nostrils in the lower frequency than 500Hz.

INTRODUCTION

Generally, sound from the nostrils is neglected except nasal sound. We have reported that sound from the nostrils is not neglected for voiced consonant[1]. This paper describes individuality for utterance of voiced consonant. From the next section, we describes the measurement of speech and analyzed results.

MEASUREMENT

Measurement is done in an anechoic room. We measured sound pressure waveform taken by a microphone, B&K 4133, at a distance of 80cm from the lips and vibration acceleration waveform on the nose taken by a small accelerometer, Endevco 2250A-10. There are 11 male subjects. Speech materials are a 16 two-syllable word list and a 6 simple sentence list in Japanese.

WAVEFORM

Figure 1 shows waveform of sound pressure, SP, (upper) and vibration acceleration, VA, (lower) when speakers A and D spoke /basu/. The beginning part of each /b/ has a buzz. When the buzz is uttered, amplitude of VA is larger than or the same as that of the following vowel /a/. From observing all data of the speech materials, it is seen that waveform of VA is not at all observed when buzz is not observed.

AMPLITUDE

To estimate amplitude of sound radiated from the nostrils for voiced consonants, we calculated as follows: (1)Analysis duration is 30msec. (2) Amplitude of VA is calculated as root mean square sample points.
Figure 2 shows the average value and standard deviation of the ratio of amplitude of voiced consonant except /m/ to /m/. It is seen that the ratio for initial consonants is larger individuality than that for second consonants.

Figure 3 shows spectra of SP and VA of /m/ when speaker A uttered /muri/. It is seen that there is much correspondence between them in less than 500Hz. Spectrum of sound radiated from nostrils can be estimated for voiced consonants, if sound is only radiated from the nostrils for nasal consonants. Figure 4 shows estimated spectra of sound from nostrils for buzz of initial consonant /b/ and vowel /a/ of /basu/. It is seen that sound is only radiated from the nostrils in the lower frequency than 500Hz for buzz of initial consonant /b/.

CONCLUSIONS

We measured sound pressure, and vibration acceleration waveform on the nose. When the buzz is uttered, amplitude of VA is larger than or the same as that of the following vowel. It is seen that waveform of VA is not at all observed when a buzz is not observed. It is seen that the amplitude of VA for initial consonants is larger individuality than that for second consonants. It is shown that sound is only radiated from the nostrils in less than 500Hz for buzz of initial consonant /b/.

REFERENCE

The Tense-Lax Question and Intraoral Air Pressure

Dae-Won Kim
Dpt. of English, Pusan National University, Pusan 609-735, S. Korea. e-m.: dwokim@hyowon.pusan.ac.kr

Measurements were made of (comparatively) flattened peak intraoral air pressure onset time (PoOT), voice cessation time (VCT), flattened peak intraoral air pressure (Po), the duration of flattened peak intraoral air pressure (drPo), pressure-fall interval (PoFI) and the duration of oral closure (DOC) as four English speakers uttered isolated nonsense V1CV2 words containing /b/ and /p/ (V1 = V2 and the V was /A/), with stress on either V1 or V2 alternately. The hypothesis tested was: The tense stop consonant will be characterized either by a higher Po or a longer drPo, and/or by both against lax. Findings: (1) PoOT was highly significantly greater in /b/ than /p/, (2) the voiceless stop /p/ produced generally greater mean Po, averaged across five tokens, than its voiced counterpart /b/, but statistically insignificant, and (3) the difference in drPo(c) between /p/ and /b/ was highly significant (p < 0.001), regardless of stress placement and subjects. The results strongly supported the linguistic hypothesis of tense-lax distinction, with /b/ being lax and /p/ tense in English. The drPoc (DOC for /p/ - PoOT for /b/) for /p/ was obtained for a reliable comparison with the drPo for /b/.

HEADING

The existing literature reveals that differences in intraoral air pressure among voiced-voiceless cognate pairs are attributed mainly to differences in glottal resistance and supra-glottal air volume, but not respiratory efforts [1, 2, 3, 4]. This claim may have something to do with air pressure at air-flow. The air pressure in /b/, measured at a point where air-flow continues, should be under the influence of glottal resistance and supraglottal air volume. The glottal resistance responsible for much of the difference in pressure is closely related to PoOT, i.e., the rising interval of pressure from the moment of oral closure to the onset of (comparatively) flattened peak intra-oral air pressure, but not Po. Voicing and some degree of glottal constriction will delay the onset of PoOT. Once intra-oral air pressure levels out, no airflow will occur, and then plateau will necessarily take place if PoOT is shorter than the duration of oral closure (DOC). The (comparatively) flattened peak intra-oral air pressure (Po) is an out-put of the equalization of air pressure between supraglottal and subglottal cavities. Thus, the Po would reflect the peak amplitude of subglottal pressure generated by the respiratory efforts during the oral closure of stop consonants, and the duration of flattened peak intra-oral air pressure (drPo) would reflect, although indirectly, the amount of energy generated by the respiratory muscle activities. Neither the Po nor the drPo has anything to do with the following phonetic events, such as the increase in supra-glottal volume [5, 6], glottal resistance, a faster post-vocalic closing movement for a higher pressure rise [7], and an active widening of the glottis for aspiration [7]. Because glottal function can not explain differences in Po and drPo among the cognate pairs, it is required to take an investigation to study pressure characteristics of stop consonants in relation to timing variables, such as PoOT, drPo, DOC and voice cessation time (VCT), in order to determine (1) whether or not Po and drPo are reliable phonetic cues for the tense-lax distinction and to see (2) how DOC is related to PoOT, Po and drPo. The hypothesis tested was as follows: The tense stop consonant will be characterized either by a higher flattened peak intra-oral air pressure or a longer duration of the flattened peak Po and/or by both against lax. In this study, the feature tensity is defined as the amount of muscle action (or energy) used in the articulation of a phoneme, and the phonetic correlates of the feature tensity are considered to be time or amplitude and/or both [8]. According to this conception, the tense stop should be characterized either by a longer respiratory activity (i.e., longer duration of flattened peak intra-oral air pressure) or by a greater peak respiratory gesture (i.e., a greater flattened peak intra-oral air pressure), and/or by both against lax.

For sensing intra-oral air pressure, a 17 cm long catheter of approximately 2 mm internal diameter was connected to the transducer and inserted into the mouth. The signals of the vocal fold vibration, air pressure and audio waveforms then were recorded simultaneously on the Siemens Oscillo-mink Mingograf running at 100 mm/s with a frequency response of 0 to 700 Hz ± 3dB and paper speed accurate to within 5%.

RESULTS AND DISCUSSION

There was a significant difference in drPo among the cognate pairs. This was attributed to differences in pressure fall interval (PoFI) and PoOT. The PoOT for /p/ was remarkably shorter than that for /b/. The mean ratio of PoOT, averaged across tokens, subjects and stress, between the cognate pairs was 2.6. The mean PoFI in /b/ was 20.7 % (27.3 ms) of DOC, while /p/ was produced with no PoFI. Po in /b/ started to fall rapidly within mean 27.3 ms (20.7 % of DOC) of release. In the production of /p/, on the other hand, the delay between lip opening and pressure drop ranged...
from 0 to 15 ms (for similar results, see [9], p. 169). Glottal openness and pressure drop lag are required to produce an acoustic cue with a puff of air, aspiration, for /p/. The aspiration is one of the principal acoustic cues to distinguish /p/ from /b/ in English.

All else being equal, the PoOT in /b/ is an output of (1) the expanded supraglottal cavity, (2) the increased glottal resistance, (3) subglottal pressure generated by respiratory muscle activities, (4) the combination of two, three or four out of them, or (5) the combination of all of them. If differences in subglottal pressure between /p/ and /b/ is insignificant [11, 12], it can be postulated that remarkably greater PoOT for /b/ than for /p/ depended largely upon differences in supraglottal air volume and the glottal resistance. In order to get reliable drPo for /p/ which is comparable with that for /b/, it is necessary to subtract the portion of PoOT in /b/ from DOC for /p/. However, it is difficult or almost impossible to calculate the exact portion of PoOT influenced only by the glottal resistance and supraglottal air volume. Thus, for the sake of convenience, the drPoc (DOC for /p/ - PoOT for /b/) for /p/ was obtained for a reliable comparison with the drPo for /b/.

The difference between the drPoc for /p/ and drPo for /b/ was highly significant (p ≤ 0.001), regardless of stress and subjects (see drP in Table 1). Considering this and the fact that Po in /p/ was generally greater than that in /b/, although statistically insignificant, it can be said with confidence at the level of 0.1% that the hypothesis that the tense stop consonant will be characterized either by a higher Po or a longer drPo, and/or by both against lax has been verified. The results of the Po and drPo strongly supported the linguistic hypothesis of tense-lax distinction, with /p/ being tense and /b/ lax. This is in contrast with the contention of a number of investigators [1, 2, 3, 4, 10] that much of differences in intra-oral air pressure among cognate pairs are attributed mainly to glottal resistance. The apparent discrepancy is due to the fact that the existing claims base their argument on air pressure under the influence of glottal resistance and we base ours largely on differences in drPo(c) which is independent of the glottal resistance. The drPo appeared to be one of convincing phonetic correlates of the feature tenseness based on respiratory efforts, although indirect, but the Po alone was insufficient to support the linguistic hypothesis of tense-lax distinction.

The shorter PoOT than DOC is always accompanied by Po and drPo. The British English subjects in this study yielded shorter PoOT than DOC in /b/, and they produced no fully voiced /b/s in isolated / CV/ words where the vowel was /A/. PoOT was composed of voicing, i.e., VCT, and devoiced period (mean 26% of PoOT). The rate of pressure rise increases during intervocalic voiced stops when voicing ceases (for similar results, see [10]). The effects of stress on Po and drPoc were inconsistent, owing to between-subject variabilities. In general, both drPo and drPoc were inconsistently related to Po.

<p>| Table 1. The results of t-test between /p/ and /b/ in PoOT, Po, drPo, DOC during isolated intervocalic bilabial stops of English (the statistical results for drPoc indicates difference between drPoc for /p/ and drPo for /b/). |
|-----------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Variabiles</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Po</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Po</td>
<td>0.12</td>
<td>1.73</td>
<td>18.5</td>
<td>0.98</td>
</tr>
<tr>
<td>drPo</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>drPo</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>DOC</td>
<td>0.25</td>
<td>0.01</td>
<td>0.13</td>
<td>0.25</td>
</tr>
<tr>
<td>PoOT</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Po</td>
<td>0.37</td>
<td>0.53</td>
<td>0.02</td>
<td>0.74</td>
</tr>
<tr>
<td>drPo</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>drPoc</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>DOC</td>
<td>0.04</td>
<td>0.12</td>
<td>0.04</td>
<td>0.12</td>
</tr>
</tbody>
</table>

REFERENCES
This study describes the vowel formant frequency characteristics of F1-F4 of Japanese five vowels produced in a fixed phonetic context. Five vowels were recorded from 28 hearing-impaired and 30 normal hearing subjects. These data obtained were analyzed by Sound Spectrograph (KAY DSP 5500). Results showed that vowel resonances of the hearing impaired were consistently lower than those of normal hearing. The range of differences between the hearing impaired and normal hearing extended about 2000Hz for F2, 3000Hz for F3 and 3500Hz for F4. The origin of vowel formant frequencies in vocal tract resonance characteristics is discussed with reference to differences in articulator behaviors.

INTRODUCTION

For many hearing-impaired speakers, improper control of respiration, phonation and articulation has negative effect on the intelligibility of speech [1,2,3]. It was the purpose of the study to analyze and compare the vowel formants of the hearing impaired and normal hearing speakers. The frequencies of first four formants of five vowels in Japanese were the main concern of this study.

METHOD

Subjects

Thirty normal-hearing subjects were drawn from university students of Tsukuba. Criteria for the selection of these subjects were no clinical speech problems and no hearing problem. Twenty-eight hearing impaired subjects were drawn from the School for the Deaf, University of Tsukuba. Criteria for the selection of these subjects were average of 90 dB or more hearing level. Hearing loss for all children was classified as severe or profound. All subjects daily use personal hearing aids.

Speech Samples

The five Japanese vowels /a/, /i/, /u/, /e/ and /o/ were selected for analysis.

Recordings

The subjects were seated in front of the microphone and the five vowels were then presented separately 18cm x 25.5cm cards. The procedure was recorded in a sound-treated recording room by means of a DAT recorder (SONY TCD-10) and the microphone (SONY ECM-959A).

Spectrographic Measurements

Digital Sonagraph (KAY DSP 5500) was used for the acoustical analysis of five vowels. This instrument has full-scale frequency range of 8000Hz. Measures were made for four lowest formants. The peak component of the vowel envelope was chosen as representative of the formant frequency.

RESULTS and DISCUSSION

The mean first four formants of the normal hearing and hearing-impaired subjects are shown in Table 1 and Table 2. These values of vowels are plotted graphically in Figure 1.

The range of the means of first formant frequency (F1) for the normal hearing are greater than for hearing-impaired subjects for vowel /a/. The range of the means of second formant frequency (F2) for hearing-impaired subjects are greater than for normal hearing for vowels /u/ and /o/. The range of the means of third
formant frequency (F3) for the normal hearing are greater than for hearing-impaired subjects for vowels /e/ and /o/. The range of the means of fourth formant frequency (F4) for hearing-impaired subjects are greater than for the normal hearing for vowel /i/.

The third and fourth formants frequency for the normal hearing are higher than for hearing-impaired subjects for five vowels.

In the selection of the formants, those of the normal-hearing speakers were located with relatively less difficulty than those of the hearing impaired. There were two reasons for this, first, the hearing-impaired subject’s production of the vowels was so distorted that a vowel identification was impossible. Second, the characteristics of hearing-impaired subjects speech itself, such as excessive aspiration, nasality, and hoarseness, interfered with clearly defined formant peaks.

Table 1. Mean first four formants of vowels for normal hearing subjects.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>527</td>
<td>249</td>
<td>220</td>
<td>1310</td>
</tr>
<tr>
<td>F2</td>
<td>1726</td>
<td>663</td>
<td>600</td>
<td>2970</td>
</tr>
<tr>
<td>F3</td>
<td>2969</td>
<td>355</td>
<td>2210</td>
<td>3870</td>
</tr>
<tr>
<td>F4</td>
<td>3703</td>
<td>235</td>
<td>3050</td>
<td>4200</td>
</tr>
</tbody>
</table>

Table 2. Mean first four formants of vowels for hearing-impaired subjects.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>500</td>
<td>155</td>
<td>240</td>
<td>960</td>
</tr>
<tr>
<td>F2</td>
<td>1525</td>
<td>405</td>
<td>740</td>
<td>2400</td>
</tr>
<tr>
<td>F3</td>
<td>2513</td>
<td>323</td>
<td>1700</td>
<td>3340</td>
</tr>
<tr>
<td>F4</td>
<td>3470</td>
<td>272</td>
<td>2660</td>
<td>3880</td>
</tr>
</tbody>
</table>

FIGURE 1. Second formant frequencies for vowels

ACKNOWLEDGMENTS

The authors are grateful for the cooperation of the following persons who arranged for the participation of the children as subjects in the study: Jiro Imai, director of the School for the Deaf, University of Tsukuba; Sawa Saito, Principal of the School for the Deaf, University of Tsukuba. I also thank Yasuko Ota, Institute of Disability Sciences, for her help.

REFERENCES