A NOTE OF DIFFERENCE LIMEN FOR ARTICULATORY PARAMETERS

Qiguang Lin and Yanli Han*
CAIP Center, Rutgers University, NJ 08855
* Qiguang Lin is now with IBM Watson Research Center, NY and
Yanli Han is with Fujitsu Networking Communications Inc, NY

ABSTRACT

This paper presents preliminary results on difference limen for articulatory parameters. An articulatory based speech synthesizer, TRACTTALK, is utilized to generate speech stimuli. Each time only one articulatory parameter is altered by a small amount. The synthetic speech is then randomized and presented to a group of listeners in the AB form. As expected, the results show that not all small changes in articulatory parameters are perceptually noticeable. The observed perceptual limen for articulatory parameters is comparable to that for acoustic parameters. The results have been used to guide the parameter optimization step in a voice mimic system. It is hoped that these results can also be useful for understanding required accuracies when coding or storing articulatory parameters, and when processing vocal-tract imaging data.

1. INTRODUCTION

Various speech processing technologies necessitate parameterization of the speech waveform. Cepstrum coefficients (including their derivatives and variants) are to date commonly used in speech and speaker recognition. Renewed efforts have recently been made to use more compact description of speech information with the help of articulatory-based speech synthesizers. Often the articulatory parameters are estimated by means of the analysis-by-synthesis technique or the voice mimicking system [1, 2, 3]. More specifically, the mimic system utilizes an articulatory-based speech synthesizer to generate synthetic speech, which moment by moment is adapted to arbitrary speech input, see Figure 1. The perceptually-weighted spectral difference between the input and synthesized speech is minimized by optimizing the underlying articulatory parameters until the difference is driven below a predetermined threshold. The optimization is in general a complex and time-consuming task because the mapping from acoustics to vocal-tract configurations is known to be not unique and nonlinear.

In order to help optimize articulatory parameters, this paper attempts to determine their perceptual limen such that incremental search steps can be reasonably chosen. First, an articulatory speech synthesizer, TRACTTALK, is described in some detail. TRACTTALK simulates the vocal tract based on principles of linear acoustics, and incorporates many known features such as wall impedance, surface losses, radiation impedance, sinus piriformis, nasal sinuses, symmetry/asymmetry of the nasal tracks, and interaction between the voice source and the vocal tract. TRACTTALK also provides parametric models for describing the vocal tract area function in terms of the minimum area, its location, the lip opening, the velum opening, the tongue tip elevation, and the overall length of the vocal tract.

Sounds are synthesized with one varying parameter at a time. In this paper, three parameters are studied: (1) the vocal tract minimum area, (2) its location, and (3) the lip opening. The resultant sounds are then randomized and presented in the AB form to a group of graduate students. The students are asked to decide if they have heard the same sound or different. The results show that the perceptual limen for a parameter is not fixed. Rather it depends upon its initial value and the value of other (fixed) parameters. For instance, the front vowels are more sensitive to change in the position of the minimum vocal tract opening than the back vowels are. When converting to formant frequency, it is found that the limen observed here is roughly comparable with that of [4].

This paper is organized as follows. A description of the articulatory speech synthesizer, TRACTTALK, is given in Section 2. Experimental results are discussed in Section 3. A summary is provided in Section 4.

2. ARTICULATORY SYNTHESIZER: TRACTTALK

Figure 1 depicts the block diagram of voice mimicking. From Figure 1, it is seen that speech can be represented with different descriptions, from text through articulation to acoustics. An articulatory description offers

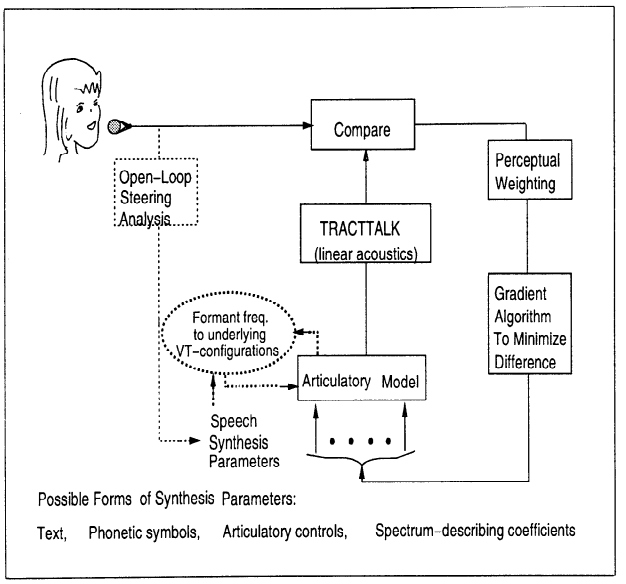

Figure 1: Block diagram of adaptive voice mimic.

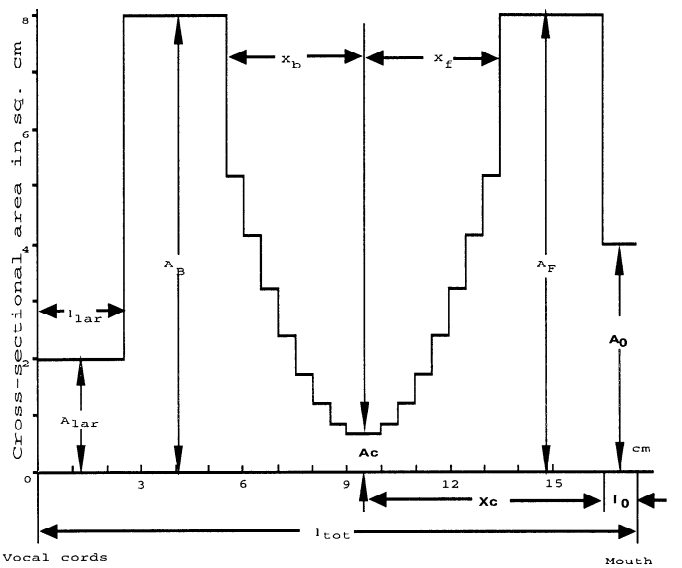

Figure 2: An example of the vocal tract area function generated by TRACTTALK.

an intermediate and effective solution. An articulation-based synthesizer has the potential to allow for a greater flexibility in adjusting for individual speaker types and speaking style, more convenience and effectiveness in specifying and interpolating control parameters, inclusion of the interaction between the voice source and the load of the vocal tract, and improved sound quality. Compared to acoustic parameters, articulatory parameters generally vary continuously at slower rates. On the other hand, a text level description is still too ambitious to handle because it further requires a complex conversion from text input to articulatory controls.

A comprehensive articulatory speech synthesizer is very important to the success of voice mimicking systems. The synthesizer we have used is the one developed at KTH and at Rutgers, TRACTTALK [5, 6, 3]. The features of TRACTTALK include:

- all major components of the speech-producing system, such as the tracheal tubes, larynx, pharyngeal cavities, oral chambers and nasal passages
- radiation impedance at the radiation ports (such as the lips, the nostrils)
- the nasal sinuses and the asymmetry of the nasal passages
- the surface losses (viscous loss and heat conduction loss)
- the wall impedance and sound radiation from the yielding walls
- for fricative sounds, the specification of the friction-source location
- inner radiation when the cross-sectional area changes abruptly between contiguous sections
- the sinus piriformis
- increase in the length of the air passage due to an extremely displaced tongue body

Furthermore, TRACTTALK includes a vocal-tract area model controlled by a few articulatory parameters (the minimum vocal-tract area, its location, the lip opening, the velum opening, the tongue tip elevation, and others). It also enables the simulation of the acoustic interaction between articulation and phonation [5].

TRACTTALK simulates the vocal tract acoustics in the frequency domain. This technique is now well established. The reader is referred to Fant [7] and Lin [5] for more details about basic principles of the vocal-tract modeling. From the frequency-domain simulation, the transfer function, $H(\omega)$, between the radiation port(s) and the exciting source(s) can be derived.

In order to synthesize speech output the transfer function, $H(\omega)$, needs to be converted to the impulse response, the time domain equivalent. There are several ways to perform the conversion, for example, the inverse Fourier transform [2]. In TRACTTALK, $H(\omega)$ is first decomposed into the zero part and the pole part. Normal modes of $H(\omega)$ are then calculated. The decomposition ensures that the calculation of poles is not contaminated by adjacent zeros, and vice versa. The next step is to determine residues at the poles and $H(\omega)$ can consequently be represented by a partial fraction expansion series. It is sufficient to only include the first few terms of the expansion within a given range of fre-

quency. Five terms are often attained for vowel sounds. Each term in the series actually corresponds to a formant generator and can be specified with a second-order digital filter [8]:

$$y_n(i) = g \cdot x(i) - a_1 \cdot x(i-1) + b_1 \cdot y_n(i-1) - b_2 \cdot y_n(i-2) \quad (1)$$

where the subscript n stands for formant index. $y(i)$ denotes the output and $x(i)$ denotes the input.

The sum of the output from individual formant generators forms the total output:

$$y(i) = \sum_{n=1}^{N} y_n(i) \quad (2)$$

or in other words, formants are connected in parallel.

3. EXPERIMENTAL RESULTS

3.1 Set-up

To determine the difference limen for articulatory parameters, TRACTTALK is used to generate speech stimuli. In this paper, only three parameters are examined for the production of vowels. They are the minimum area of the vocal tract (A_c), its location (X_c), and the lip opening (L_0/A_0). These three parameters are defined in Figure 2. Other parameters are introduced to cope with more complex vocal tract configurations and are usually set to default constants. The overall vocal tract length in the following experiments is set to 17.5 cm.

Two types of glottal excitation are provided in TRACTTALK. One is the LF model, and it is noninteractive. The other type is an interactive glottal source which takes into account the nonlinear effects between the vocal tract load and the glottal source. In the experiments described below, the noninteractive LF model is used as the input x of Eq. (1). The values of LF parameters are taken from the result of inverse filtering an adult male speech. It should be noted that the LF model simulates the derivative of the glottal flow, and hence already incorporates a differentiator.

Several anchor positions are chosen, and the three articulatory parameters are perturbed in the vicinity of the anchors. In this paper, we only consider three anchors of $X_c = 4, 8$, and 11 cm. A_c and l_0/A_0 are each fixed to two values to produce three combinations. X_c is slightly altered in a step of 0.3 cm to cover a dynamic range of about ± 2.0 *cm* from the anchor.

Synthetic speech for each chosen vocal tract configuration is then generated. The stimuli are randomized and presented in the AB form to a group of graduate students (7 males and 5 females). The subjects have to make a forced decision whether a pair of two vowels sound the same or differently.

3.2 Results

The listening results are plotted in Figure 3. The solid curves are for female, the dotted curves are for male, and the dashed curves are for the pooled results. It is seen that in most cases the difference between the three curves are small. 0 denotes that no subjects discriminate the paired sound stimuli, and 100% denotes of course that all subjects discriminate the stimuli.

Figure 3 shows that the back vowels (larger X_c values) are less sensitive to changes in X_c than the front vowels. The sensitivity of front vowels increases slightly when A_c decreases. The increased sensitivity at $X_c = 4$ *cm* may be due to the fact that it is next to the location of the maximum F_2 and F_3 proximity [7]. Around this location both F_2 and F_3 change substantially as a function of X_c. The smaller A_c is, the more apparent change in F_2 and F_3. As a result, a slight change in X_c can be more easily detected. In addition, increased lip rounding also causes F_2 and F_3 to change more substantially as X_c varies around the proximity. This increased discriminability is clearly seen in the middle panel. It is also noted that the cavity-mode affiliation changes when passing the maximum F_2 and F_3 proximity. Such changes may have additionally contributed to the increased discriminability for $X_c = 4$ *cm* because of changes in bandwidths.

Furthermore, Figure 3 shows that in all cases examined human speakers are not able to to discriminate a change of ± 0.3 *cm* in X_c. For $X_c = 8$ *cm* and $X_c = 11$ *cm*, a change of ± 0.6 *cm* in X_c does not warrant a noticeable difference perceptually.

We have also examined corresponding changes in formant frequencies (not shown here). It is found that the results are roughly comparable to the just noticeable difference for the acoustic parameters reported in [4].

4. SUMMARY

In the above, we have described an articulatory speech synthesizer, TRACTTALK. The synthesizer is used to generate speech stimuli to experimentally determine the difference limen for articulatory parameters. We have shown discrimination curves for vowel-likes configurations for three anchors of X_c under three combinations of A_c and l_0/A_0. The results have been reported and discussed in Section 3.2.

It is important to note that this study uses sustained sounds. It has been shown, in the case of pitch perception, the differential threshold increases in a dynamic situation [9]. Use of an interactive glottal source may also modify the discrimination curves

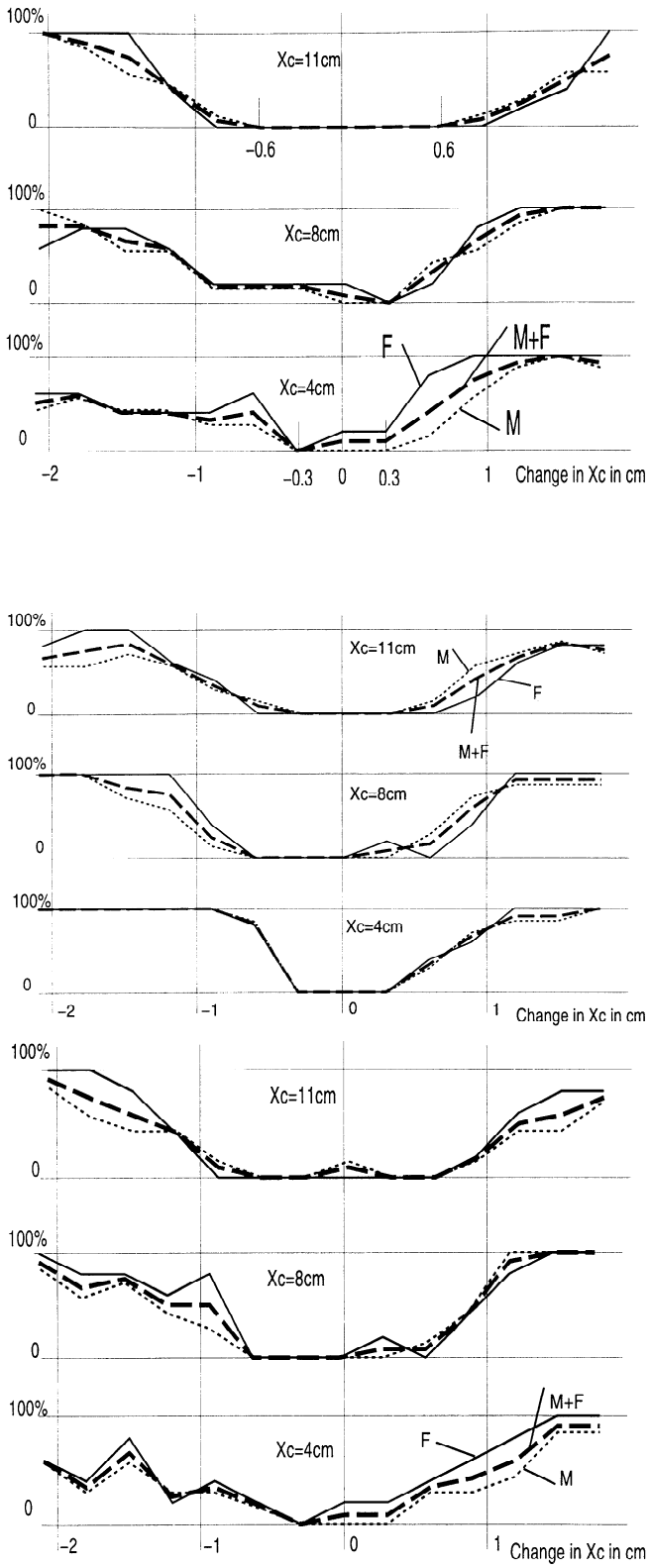

Figure 3: Just noticeable difference for X_c. Top: $A_c = 0.65\ cm^2$ and $l_0/A_0 = 0.25\ cm^{-1}$; Middle: $A_c = 0.65\ cm^2$ and $l_0/A_0 = 1.0\ cm^{-1}$; Bottom: $A_c = 1.2\ cm^2$ and $l_0/A_0 = 0.25\ cm^{-1}$.

slightly. Nonetheless, sustained sounds establish the lower bound of just noticeable difference and there is no need to require a description accuracy or "fidelity criteria" below that bound. The described experiments can be extended to other dimensions of the articulatory space.

We have successfully used the presented results to guide the design of the codebook relating articulatory parameters and acoustics characteristics. The codebook is for selection of an initial vocal tract shape for iterative voice mimicking. We have also used the results to guide the optimization step of our voice mimic system. It is hoped that the results can be utilized to interprete the quantal theory of speech production [10].

Acknowledgments

This work is a component of the High Performance Computing and Design (HPCD) project, and has been supported in part by an ARPA contract #DAST-63-93-C-0064. The authors whish to thank J.Y. Zou for his assistance and the 12 subjects. The TRACTTALK program is available for research purposes upon request to qlin@us.ibm.com.

References

[1] J. Flanagan, K. Ishizaka, and K. Shipley, "Signal models for low bit-rate coding of speech," *J. Acous. Soc. Am. 68*, pp. 780-791, 1980.

[2] J. N. Larar, J. Schroeter, and M. M. Sondhi, "Vector quantization of the articulatory space," *IEEE-ASSP 36*, pp. 1812-1818, 1988.

[3] Q. Lin, G. Richard, J. Zou, D. Sinder, and J. Flanagan, "Use of TRACTTALK for adaptive voice mimic," *J. Acous. Soc. Am. Vol. 97* (abstract), 1995.

[4] J. Flanagan, "A difference limen for vowel formant frequency," *J. Acous. Soc. Am. 27*, pp. 613-617, 1955.

[5] Q. Lin, *Speech production theory and articulatory speech synthesis*, Ph.D. thesis, Royal Institute of Technology (KTH), Stockholm, 1990.

[6] Q. Lin and G. Fant, " An articulatory speech synthesizer based on a frequency-domain simulation of the vocal tract," *ICASSP*, San Francisco, 1992.

[7] G. Fant, *Acoustic Theory of Speech Production,* (Mouton, the Hague), 1960.

[8] Q. Lin, "From vocal-tract transfer function to the impulse response," *IEEE-Trans Speech & Audio Processing*, Vol. 3, pp. 449-457, 1995.

[9] D. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acous. Soc. Am. 67*, pp. 971-995, 1980.

[10] K. Stevens, "The quantal nature of speech: Evidence from articulatory-acoustic data," In *Human Communication: A Unified View* (eds. David and Denes), pp. 51-66, 1972.