SPEECH SYNTHESIS FROM REAL TIME ULTRASOUND IMAGES OF THE TONGUE

Bruce Denby† and Maureen Stone‡

†Laboratoire des Instruments et Systèmes d'Ile de France, Université Pierre et Marie Curie, B.C. 252, 4 place Jussieu, 75252 Paris Cedex 05, France ; denby@ieee.org
‡Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore, MD, 21201

ABSTRACT

A machine learning technique is used to match reconstructed tongue contours in 30 frame per second ultrasound images to speaker vocal tract parameters obtained from a synchronized audio track. Speech synthesized using the learned parameters and noise as an activation function displays many of the time and frequency domain characteristics of the original audio, and for isolated passages, is remarkably clear – although no articulators other than the tongue are included.

1. INTRODUCTION

Although medical ultrasound has been used in articulatory speech modelling research for many years [1,2,3], a voice interface based on the technique would until now have been considered far too cumbersome. Advances in microelectronics, however, are changing this scenario. Compact, real time interfaces using non-invasively sensed articulatory cues could be used, for example, to enhance speech quality in noisy environments; to directly synthesize digital speech when audio silence is required; or in prosthetic applications for handicapped individuals.

This article presents a first attempt to develop a speech synthesizer driven by tongue contours acquired with a real time medical ultrasound [4]. Extracted contours, mapped onto vocal tract parameters and combined with noise, yield a speech signal which already reproduces many of the spectral and temporal characteristics of the original recording, and in certain passages is of remarkable fidelity. Experimental details, data analysis techniques, a discussion of the results, and future prospects are presented in the following sections.

2. DATA ACQUISITION

Data were taken on an Acoustic Imaging Performa 30 Hz ultrasound machine [5] with a 2-4 MHz, 96 element curvilinear array, using the HATS system [6] to immobilize the speaker's head and support the transducer under the chin in a fixed position. The speech corpus consisted of a 6-sentence passage called *Rainbow* and a 9-sentence one called *Grandfather*, designed to contain multiple examples of all English phonemes. A native English speaker repeated each passage twice, creating a total of 149.7 seconds of speech stored in 4491 .jpg ultrasound frames and one 11025 Hz-sampled time-synchronized .wav audio file for each of the 30 sentences.

3. EXTRACTION OF TONGUE CONTOURS

The ultrasound images were first reduced to a 14 (radial, or time axis) by 40 (azimuthal angle) grid, superimposed on the original fan-shaped data field and enclosing all possible tongue contour points of interest. (The grid simulates the readout one might obtain from a simple, 14 element probe – a point returned to in section 7). Candidate contour points in each time bin were attributed using a maximum smoothed spatial intensity gradient criterion. Non-bifurcating contours were then constructed by linking candidate points *via* a local smoothing algorithm [7]. A final filtering step corrected contours by requiring smoothness between consecutive frames in time as well. A typical result is shown in figure 1, where an 'r' is pronounced.

4. LEARNING VOCAL TRACT PARAMETERS

The vocal tract model of the GSM vocoder [8] was chosen because of its readily available code and proven ability for real time implementation - an important consideration in this study. The 13 kbit/sec GSM codec transforms blocks of 160, 13-bit speech samples (20 ms at 8000 samples/second) into 260 bits of coded information as outlined in table I. A machine learning algorithm was used to map the 14 tongue contour points onto a subset of the codec parameters in each block.

To simplify the data handling, the .wav audio files were downsampled to 160 samples per ultrasound frame (i.e., 4800 samples/sec), so that the GSM codec could output exactly one speech block per frame. This is of course a compromise, as 1) undersampling the audio from the original 11025 Hz causes some degradation of the signal, and 2) the signal will seem 'speeded up' to the GSM codec, which expects an 8 kHz sampling frequency.

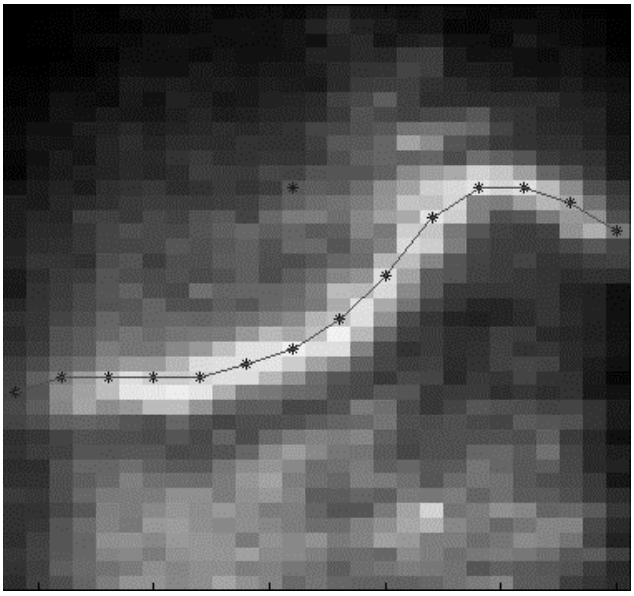

Figure 1. Fiducial region of ultrasound image containing all tongue contours. Vertical axis is time, horizontal angle. Only every other angle bin is used, giving a 14 by 40 grid. Candidate contour points are shown as stars, final non-bifurcating tongue contour as solid line. Here, an 'r' is pronounced. Tongue tip is at right.

Tongue contours should in principle only provide information on the 8 log-A ratios of the LPC filter. However, the machine learning algorithm was extended also to the 20 ms averages of the LTP *lag* and *gain* (≈ pitch period) and RPE (activation function) *ssseq* and *xmax* variables, since, empirically, statistical correlations between these additional variables and the LPC ones were found to exist in the data. Only the 13 RPE residual samples were entered as noise, leaving 12 GSM variables to be learned from the tongue contours.

The algorithm chosen was a 14-20-20-12 multilayer perceptron (MLP) trained with gradient backpropagation [9]. Ninety percent of the data was used for training and 10% to assure that overtraining of network weights did not occur. Variant MLP architectures, as well as tests with different subsets of output variables, gave similar results. Once learned, the GSM parameters and noise activation function were used to write GSM audio files which were then expanded, using the standard GSM decoder, into output .wav files. These files incorporated training plus test data; spot checks indicated that doing so introduced no biases.

5. EVALUATION OF RESULTS

Listened to by themselves, the synthesized passages are not recognizable; however, by alternating between original and synthesized recordings, the listener can soon easily pick out most of the correspondences. As the quality is insufficient to allow the use of a measure such as PESQ [10], it is more instructive here to examine the overall temporal and frequential aspects of the signal and to give a qualitative idea of the performance of the method on different utterances.

RPE-LTP GSM Vocoder		**bits/5 ms**	**bits/20 ms**
LPC filter	8 log-A ratios	-	36
LTP filter	lag	7	28
	gain	2	8
RPE activation function	ssseq	2	8
	xmax	6	24
	13 resid. samp.	39	156
total			260

Table I. Contents of a 260-bit, 20 ms GSM speech block.

Figure 2 presents the spectrogram of a sample sentence from the *Grandfather* passage. One may make the following observations:

- The learning algorithm is, with rare exceptions, able to distinguish between speech and silence. This is an important result which was far from obvious at the outset of this study.
- The envelopes of the frequency spectra of the synthesized utterances are in rough accord with those of the original passage. Since the activation function is noise, one does not expect, of course, to see in the synthesized speech the banded harmonic structure evident in the original signal.

Also shown in figure 2 is a decomposition of the sentence into 5 segments labeled *a* to *e*. The content of these segments is given below, where those utterances which seem the most clear are printed in underlined capitals:

a) wE have Often Urged hIm to …
b) … WALK mORE …
c) … and smOKE LESS …
d) … but he ALways ANswERS …
e) … bANANA OIL …

Thus, the MLP appears to have learned best the tongue contours which are the most distinctive - 'l', 'r', 'k', as well as many of the vowels - but does less well on the phonemes for which it has little information, such as plosives and fricatives. As a check of reproducibility, it should be noted that the breakdown into 'good' and 'incorrect' segments is nearly identical in the two repetitions of each passage.

One should also expect that certain phonemes *cannot* be uniquely identified from only the rather crudely measured tongue contours used in this study. This corresponds to a situation in which the MLP is requested to map nearly identical inputs onto different outputs. Since backpropagation is a least-squared error minimization algorithm, the MLP will in these cases give output variables near the centroids of those of the ambiguous output classes. This is consistent with the observation that 'incorrect' synthesized utterances, rather than being wildly inappropriate, tend for the most part to have a neutral, droning quality.

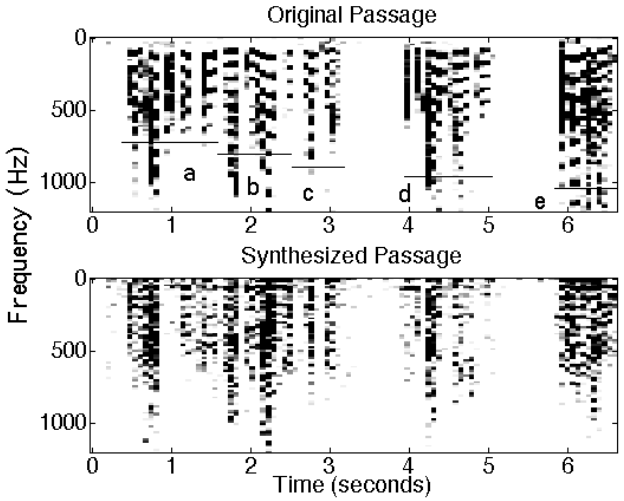

Figure 2. Spectrogram of a sample sentence from the *Grandfather* passage. The segments a-e are described in the text.

6. FUTURE IMPROVEMENTS

A number of possible future improvements seem clear:

- **Choice of Vocoder**. The GSM algorithm has been useful for a first study, but in the long term it will be preferable to match the vocoder used to the ultrasound frame rate. This, in particular, will allow to sample the audio at a more acceptable frequency than the 4800 samples/sec used thus far.
- **Improved Contour Finding**. Hand scanning reveals the algorithm occasionally misses a point or two at the base or tip of the tongue. Better performance, particularly in the tip region, could have a significant effect.
- **Study of Activation Functions.** The results described used noise as an activation function. Additional tests using a uniform *voiced* activation function improved some utterances but degraded others. It would be interesting to use different mixes of voiced and unvoiced activations to see if a better overall result can be extracted.
- **Trajectory Modelling.** No attempt was made in the presented results to ensure that the articulatory trajectories obtained were physically reasonable and smooth; the system simply outputs one GSM block per ultrasound frame. Use of trajectory modelling [11] would almost undoubtedly correct a number of poorly reconstructed frames.
- **Incorporation of other Articulatory Cues.** A video frame of the speaker's lips was embedded in each ultrasound frame. Lip information included with the tongue contour would quite likely disambiguate the output classes of certain utterances. Too, there is occasionally other viable information in the ultrasound image besides the tongue contour, corresponding to lip movement, hyoid bone, teeth, muscles, etc., which might be exploitable.
- **Real Time Feedback.** A speaker trying to make himself understood in a noisy or distorting environment modifies his delivery by speaking more slowly and enunciating more clearly. This possibility is lost when pre-recorded input files are used. It seems quite likely that a speaker having real time feedback of how his voice sounds at the receiving end could greatly improve the 'signal to noise ratio' over that observed in our tests.

7. REAL TIME IMPLEMENTATION

Beyond the potential for operator feedback, a real time implementation – and one that can be made compact and portable – is essential if an ultrasound speech synthesizer is to become a workable tool. Although the analyses presented here were done offline, contour finding and MLP execution can be accomplished in a only few milliseconds on a standard PC. Image acquisition is already fast enough, and the GSM algorithm runs handily on millions of cellphones every day. The required real time implementation thus seems assured.

The issue of portability also appears to be well in hand. Several ultrasound firms already offer PC-based machines [12], or even complete ultrasound systems packaged as peripherals to be connected to a standard PC [13]. Our speech synthesizer, however, does not need to be as complicated as a full ultrasound system, nor does it require an entire PC to function. We have already demonstrated that a simple ultrasound probe with 14 elements would be sufficient, and a study [14] has shown that such a probe equipped with a readout system based on Field Programmable Gate Arrays could indeed be built as a portable, handheld real time device.

8. CONCLUSION

Experimental tests of speech synthesis from real time ultrasound images of the tongue have been presented. The resulting speech is as yet of poor quality, but has many of the desired properties and can hopefully be brought to a usable level with some of the suggested improvements. If so, it should be possible, using standard technology, to realize a portable ultrasound speech interface, which could be used to enhance standard speech in hostile environments, produce speech output without glottal activity, or post-treat the speech of certain handicapped individuals to improve intelligibility.

9. ACKNOWLEDGEMENTS

The authors wish to thank the reviewers for valuable comments and suggestions for improving this paper.

10. REFERENCES

[1] B. Sonies, T. Shawker, T. Hall, L. Gerber, and S. Leighton, "Ultrasonic Visualizaion of Tongue Motion During Speech," *J. Acoust. Soc. Am,* **70**, pp. 683-686, 1981.

[2] E. Keller and D. Ostry, “Computerized Measurement of Tongue Dorsum Movement with Pulsed Echo Ultrasound,” *J. Acoust. Soc. Am.* **73**, pp. 1309-1315, 1983.

[3] M. Stone, B. Sonies, T. Shawker, G. Weiss, and L. Nadel, “Analysis of Real-Time Ultrasound Images of Tongue Configurations Using a Grid-Digitizing System,” *J. Acoust. Soc. Am.***11**, pp. 207-218, 1983.

[4] A body of work on MRI based speech synthesis exists; see for example, Olov Engwall, "Synthesizing static vowels and dynamic sounds using a 3D vocal tract model," 4th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW4), Perthshire, Scotland, August 29-September 1 , 2001, and references therein.

[5] Acoustic Imaging Technologies Corporation, Phoenix, Arizona.

[6] M. Stone and E. P. Davis, “A Head and Transducer Support System for Making Ultrasound Images of Tongue/Jaw Movement,” *J. Acoust. Soc. Am.* **98**, pp. 3107-3112, Dec. 1995.

[7] B. Denby, “Neural Networks and Cellular Automata in Experimental High Energy Physics,” *Comp. Phys. Comm.* **57** pp. 429-448, 1988.

[8] “Digital Cellular Telecommunications System (Phase 2+); Full-Rate Speech Transcoding (3GPP TS 46.010 version 5.0.0 Release 5),” ETSI TS 146 010 V5.0.0 (2002-06), ETSI**,** 650 Route des Lucioles, F-06921 Sophia Antipolis Cedex, France, available from http://www.etsi..org.

[9] S. Haykin, *Neural Networks, a Comprehensive Foundation*, Prentice Hall, New Jersey, 1994.

[10] “Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs,” ITU-T Recommendation P.862, ITU, Geneva, Switzerland, 2001.

[11] M.J. Russell, P.J.B. Jackson, and M.L.P. Wong, “Development of articulatory-based multi-level segmental HMMs for phonetic classification in ASR,” in *Proceedings of EURASIP Conference on Video/Image Processing and Multimedia Communications*, EC-VIP-MC~2003, Zagreb, Croatia, June, 2003.

[12] Online, http://www.aloka.com
Online, http://www.gemedicalsystems.com

[13] Online, http://www.terason.com

[14] P. Schweitzer, E. Tisserand, B. Denby, “Design of an Ultrasonic Lingual Profilograph,” World Congress on Ultrasonics, Paris, France, .September 7-10, 2003, to appear in the proceedings.