PROSPECTS FOR A SILENT SPEECH INTERFACE USING ULTRASOUND IMAGING

Bruce Denby[1], Yacine Oussar[2], Gérard Dreyfus[2], Maureen Stone[3]

[1]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
[2]Laboratoire d'Electronique, Ecole Supérieure de Physique et Chimie Industrielles de la Ville de Paris (ESPCI-Paristech),
10 rue Vauquelin, 75231 Paris Cedex 05, France
[3]Vocal Tract Visualization Lab, University of Maryland Dental School,
666 W. Baltimore Street, Baltimore, MD, 21201, USA

ABSTRACT

The feasibility of a silent speech interface using ultrasound (US) imaging and lip profile video is investigated by examining the quality of line spectral frequencies (LSF) derived from the image sequences. It is found that the data do not at present allow reliable identification of silences and fricatives, but that LSF's recovered from vocalized passages are compatible with the synthesis of intelligible speech.

1. INTRODUCTION

There has been interest recently in the idea of a sensor-based system allowing speech communication via the standard articulators, but without glottal activity – that is, a silent speech interface. Possible applications include a silent cellphone, silent voice data entry system, or an alternative to tracheo-oesophagal speech (TES) for persons having undergone a tracheotomy. Approaches using ultrasound imaging [1] and electromyography [2] have appeared in the literature. X-rays and magnetic resonance imaging (MRI) [3-5], though of excellent spatial resolution, are probably not applicable here due to health and portability issues. This article addresses the viability of the ultrasound option by evaluating the quality of the imagery-extracted phonetic parameters using spectral distortion measurements and informal listening tests.

The work is based on an ultrasound dataset with a lip profile image embedded in each frame, along with a synchronized audio track. Section 2 details data acquisition and preprocessing, while the machine learning approach used to map tongue and lip contours onto LSF's is described in section 3. Problems encountered in an initial analysis pass – due to ambiguities between vocalized and unvocalized phones – are discussed in section 4, and some interpretative commentary given. Spectral distortion measurements and informal listening scores on voiced speech – the principal focus of this article – are presented in section 5. The article closes with conclusions and perspectives for the future.

2. DATA ACQUISITION AND PREPROCESSING

Data were taken using an Acoustic Imaging Performa 30 Hz ultrasound machine [6] with a 2 to 4 MHz, 96 element curvilinear array. The University of Maryland HATS system [7] was employed to immobilize the speaker's head and support the transducer beneath the chin (ultimately, of course, a lighter, wearable system is envisaged). An example image is shown in figure 1.

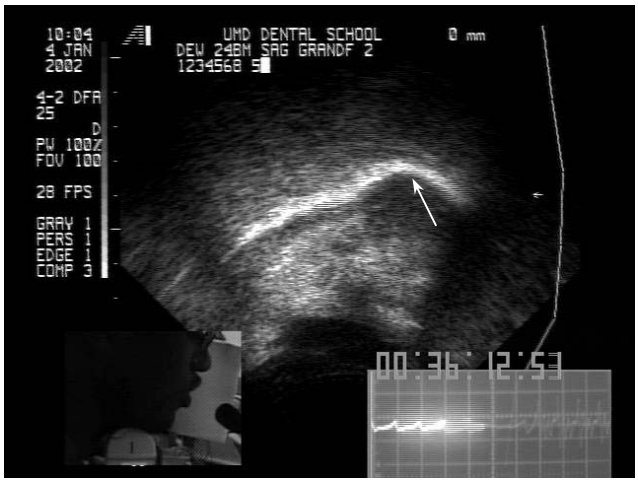

Figure 1. Example ultrasound image showing tongue contour (arrow; tongue tip is to the right) and embedded lip profile image (the insert at the lower left of the image).

Tongue contours were extracted using a maximum smoothed spatial intensity gradient criterion, and were stored as the *r* values measured at 14 equally spaced fixed θ points (*r=0* is at the center of the US probe). The lip contours were obtained by simple binarization of the profile image, and the *x-y* positions of the horizontal extrema of upper and lower lips, as well as that of the lip commissure, were then stored. The *x-y* coordinates of points a small distance above and below each of these points were also stored, in order to provide indicators of lip rounding and lip opening angle. The input to the machine learning algorithm

thus consisted of the 14 tongue *r* values plus 9 *x*-*y* pairs for the lips, for a total of 32 inputs, as shown schematically in figure 2.

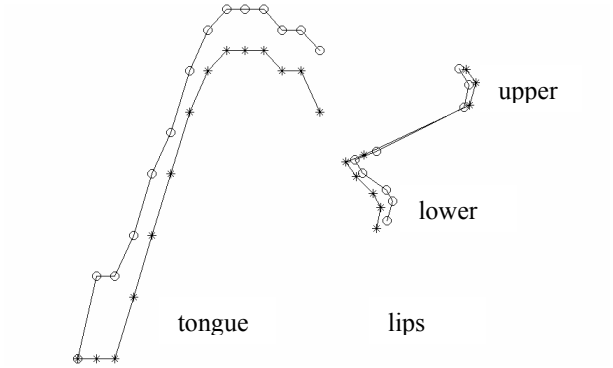

Figure 2. Data from 2 frames, one marked by circles, the other by asterisks (tongue-lip distances are not to scale). To the left, the 14 tongue contour *r* values; at right, the 9 lip contour *x*-*y* values.

The speech corpus consisted of phonetically balanced 6-sentence *Rainbow* and 9-sentence *Grandfather* passages, each repeated twice, for a total of 30 sentences. The resulting 149.7 seconds of speech was stored as 4491 .jpg ultrasound frames and 30 11025 Hz .wav audio files. LSF-based synthesis is known to be more robust against distortion compared to using, for example, LPC coefficients [8]. Twelve target LSF's were calculated for each 33.3 ms frame using linear predictive coding and a hanning window with a symmetric half-frame overlap. The residual signal from each frame was also retained.

The speech corpus is not large enough to warrant more aggressive modelling using, for example, Hidden Markov Models. The focus of this article is to evaluate the capacity of the images to furnish viable phonetic information on a frame by frame basis. A larger corpus is under study.

3. MACHINE LEARNING ALGORITHM

Multilayer perceptrons (MLP) [9] were used to perform the mapping between the 32 input variables and the 12 LSF's. A separate network was used for each LSF, rather than a single fully connected net, in order to reduce the number of adjustable parameters in the model. Before training, a variable selection procedure [10] removed between 1 and 5 of the least salient inputs from each LSF network. Thirty-one outlier frames in which the automatic contour finding had failed were removed from the training set. The training minimized a weighted least squares cost function given by

$$J = \frac{1}{2} e W e^T ,$$

where e is the vector of LSF errors, and the matrix of weights W,

$$w_{ij} = \frac{1}{LSF_{ij} - LSF_{(i-1)j}} + \frac{1}{LSF_{(i+1)j} - LSF_{ij}} ,$$

(i is the LSF index, j the frame number) originally introduced by Laroia et al. [11] for weighting spectral distortion measures, favors examples in which the LSF being trained is near a formant. This gave a small improvement in performance over an unweighted error.

Following the methodology used in [12], model selection was performed with the the *virtual leave one out* [13] (also called *PRESS statistic* [14]) method, since it allows to use the entire data set for training (earlier tests with 90% train and 10% test gave similar results). Models of increasing complexity were trained, and the best model for each LSF retained. Typically, the selected networks contained fewer than 5 hidden units.

4. DISAMBIGUATION

A first training pass revealed that the system was unable to faithfully reproduce the larger excursions of the LSF values, remaining instead in mid-range, nearer the mean. To explore the problem, a k-means algorithm [15] automatically clustered the data into 150 classes of tongue/lip contours, and the LSF vectors associated with each class examined. It was discovered that many input contour classes contained two, or even three clusters of LSF vectors, corresponding to voiced speech, silences, and/or fricatives. The learning algorithm in those cases had simply learned the mean of the often rather diverse LSF vectors present in the class. A contour class containing all three types of LSF is shown in figure 3.

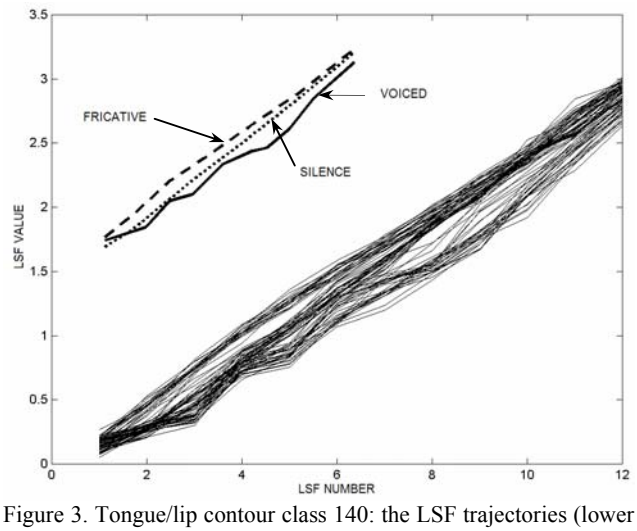

Figure 3. Tongue/lip contour class 140: the LSF trajectories (lower part of figure) form 3 clusters. The pictogram (upper part) identifies these as voiced, silent, and fricative LSF trajectories. The algorithm learns the average of the 3 classes, leading to poorer performance.

That longer, intersyllabic silences exhibit ambiguity is not surprising, as the tongue and lips need not be in any particular position during such intervals (an unambiguous "rest" position of the tongue in very long silences was, however, observed). The implication is that users of a silent speech interface will have to learn to use some mechanism other than their glottis, presumably supplied by the interface, to control the excitation of their speech waveforms, much as is the case today for users of TES or electrolarynxes.

What is more troublesome is ambiguity between voiced sounds and stops or fricatives, which are crucial to the production of intelligible speech. This result is unexpected, as stops/fricatives should in principle correspond to rather well defined configurations of the vocal apparatus. As the size of the corpus is not sufficient to study the phenomenon in detail, the decision was made to simply remove all silences, stops, and fricatives from the training set – leaving a total of 2559 voiced frames – and to concentrate upon the ability of the system to learn voiced speech. The selection of frames to remove was based on their Euclidean LSF distance from average silent and fricative frames. It is hoped that with a larger training set, more sophisticated image processing, and in the future the inclusion of additional sensors, the disambiguation of stops and fricatives will become possible.

5. QUALITY ASSESSMENT

The result of the training on voiced speech is shown in figure 4. LSF's 2, and 4-8 appear to be the easiest to learn from the tongue and lip images. For the remaining LSF's, essentially just the mean was learned.

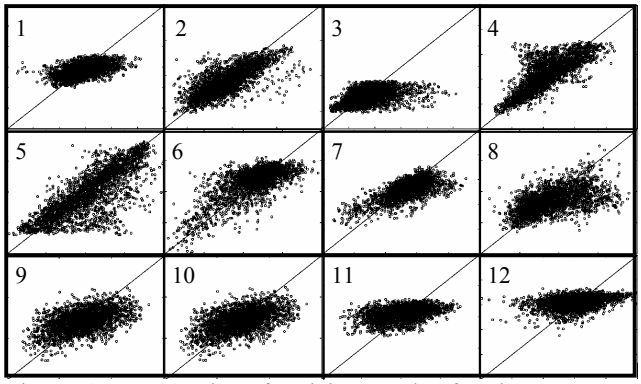

Figure 4. Scatter plot of training results for the 12 LSF's. Horizontal axis: true LSF; vertically, learned LSF.

The critical issue for a silent speech interface is intelligibility, which can only be evaluated using subjective listening tests carried out on synthesized passages. High quality synthesis will only be possible with a larger training corpus which allows the use of phonetic trajectory modelling. In this article, three simpler tests are employed in order to give some idea of the degree of intelligibility that one might expect in the final system.

The tests chosen were: the mean RMS log spectral distortion, SD, due to the imperfect learning of the LSF's; a differential mean opinion score (MOS) based on SD; and informal listening tests on LPC-reconstructed speech using residual and noise activations. The spectral distortion in dB is calculated in the classical way using:

$$SD = \left\langle \sqrt{\frac{1}{n_1 - n_0}\sum_{k=n_0}^{n_1-1}\left(10\log_{10}\left(\left|\frac{A\left(e^{j2\pi k/N}\right)}{A'\left(e^{j2\pi k/N}\right)}\right|^2\right)\right)^2} \right\rangle$$

where A, A' are, respectively, the LPC polynomials derived from the original LSF's and the learned LSF's; and N = 512, n_0 = 6, n_1 = 200, giving a frequency range of 129 – 4307 Hz and a bin size of 21.5 Hz. The differential MOS score is evaluated with respect to a "transparent" 1 dB distortion using the empirical relation [16]:

$$MOS = 3.56 - 0.8 \cdot SD + 0.04 \cdot SD^2$$
$$\Delta MOS(SD) = MOS(SD) - MOS(1\ dB)$$

For comparison, SD and MOS values were also calculated at two additional points: one from an early trial in which silences and fricatives were retained during training (but not in evaluating SD), and another using the fixed, mean values of the true LSF's. The informal listening tests consisted of having a few colleagues listen to the synthesis results and give their impressions. As the training did not produce LSF's for silences and fricatives, artificial values were used, consisting of the mean LSF vector over silent frames for the silences (coupled with a factor of 2 reduction in amplitude), and a random choice of 5 fricative LSF vectors taken from the original training set. Results are summarized in Table I. In the last line of the table appears the comparison made using the true original LSF values (artificial silences and fricatives were not used in line 5a, in order to have one "perfect" file). Files used in the listening tests (lines 1, 3, and 5) may be consulted online [17].

The results show that the machine learning substantially improved both numerical performance and listening intelligibility as compared to using the means, and that removing silences and fricatives further improved the results on voiced frames. The listening test for line 2, though not included in the table/website due to differing conditions, gave results slightly worse than those of line 3. At SD = 4dB, the learned vocal LSF's are still far from "transparent," but, if ΔMOS is any measure, should not lead to catastrophically lower intelligibility. This notion is supported by the listening tests, which suggest that from a perceptual standpoint, the learned LSF's are almost as good as the original ones, at least for this type of test. In

particular, using the learned LSF's with the true residual gave very acceptable speech (line 3a). Of course, in a real silent speech interface, one will not have the residual, and the results obtained here using a noise activation are probably not yet good enough to be usable. There is also still the issue of the silences and fricatives. A more elaborate synthesis test on a larger corpus is being developed.

Table I. Spectral distortion SD on voiced frames, ΔMOS, and informal listening test results. ΔMOS is measured with respect to a 1 dB "transparent" benchmark. Artificial silences and fricatives were used on lines 1, 3, and 5b, as explained in the text. Files used in the line 1, 3, and 5 listening tests are consultable online [17]. (The listening test for line 2, not included on the website due to differing conditions, gave results somewhat worse than line 3.)

#	Method	SD dB	Δ MOS	Listening Test	
				Activ.	Comments
1	mean LSF's	5.7	–2.5	a: resid	very distorted
				b: noise	modulated noise
2	LSF's learned on all frames	4.9	–2.2	–	
3	LSF's learned on voiced frames	4.0	–1.8	a: resid	a bit worse than 5a
				b: noise	a bit worse than 5b
4	"transparent"	1.0	0.0	–	
5	true LSF's on all frames	0.0	+0.76	a: resid	perfect
				b: noise	whispery; fair intelligibility

6. CONCLUSIONS AND PERSPECTIVES

It has been shown that sagittal ultrasound tongue contours and lip profile information are not at present sufficient for learning the line spectral frequencies of silent and fricative speech frames. On voiced speech, however, the machine learning results seem very promising, from a spectral distortion and informal listening test viewpoint. If disambiguation of silent and fricative frames can be achieved, via a larger training corpus and more sophisticated image and speech processing tools, it thus seems likely that a real time silent speech interface based on ultrasound and lip video will be feasible. Work on a much larger corpus is currently underway.

7. ACKNOWLEDGEMENTS

The authors acknowledge useful contributions from F. Berthommier, M. Milgram, G. Chollet, and the reviewers.

8. REFERENCES

[1] B. Denby and M. Stone, "Speech Synthesis from Real Time Ultrasound Imagery of the Tongue," *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP04*, Montréal, Canada, pp. 685-688, May 17-21, 2004.

[2] C. Jorgensen, D.D. Lee, S. Agabon, "Sub Auditory Speech Recognition Based on EMG/EPG Signals," *Proceedings of the International Joint Conference on Neural Networks*, vol. 4, pp. 3128-3133, Portland, Oregon, July 20 - 24, 2003.

[3] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, C. Savariaux, "Three-dimensional Linear Articulatory Modeling of Tongue, Lips and Face, Based on MRI and Video Images," *Journal of Phonetics* **30**, 533–553, 2002.

[4] P. Badin and C. Abry, "Articulatory Synthesis from X-Rays and Inversion for an Adaptive Speech Robot," *Proceedings of the International Conference on Spoken Language Processing ICSLP '96*, Philadelphia, PA, pp. 1125-1128, October 3-6, 1996.

[5] O. Engwall, "Synthesizing static vowels and dynamic sounds using a 3D vocal tract model," *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW4)*, Perthshire, Scotland, August 29-September 1, 2001.

[6] Acoustic Imaging Technologies Corp., Phoenix, Arizona.

[7] 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.

[8] G. Kang, L. Fransen, "Application of Line-Spectrum Pairs to Low-Bit-Rate Speech Encoders," *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP85*, Tampa, Florida, pp. 244- 247, March, 1985.

[9] K. Hornik, M. Stinchcombe, H. White, "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks* **2**, pp. 359-366, 1989.

[10] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a Random Feature for Variable and Feature Selection," *Journal of Machine Learning Research* **3**, pp. 1399-1414, 2003.

[11] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and Efficient Quantization of Speech LSP Parameters using Structured Vector Quantizers," *Proc. IEEE Int. Conf. On Acoustics, Speech, Signal Processing*, Toronto, Canada, pp. 641–644, May 1991.

[12] G. Dreyfus, *Neural Networks, Methodology and Applications*, Springer, 2005.

[13] G. Monari, G. Dreyfus, "Local Overfitting Control via Leverages," *Neural Computation* **14**, pp. 1481-1506, June 2002.

[14] S. Chen, X. Hong, C. J. Harris, P. M. Sharkey, "Sparse Modeling Using Orthogonal Forward Regression With PRESS Statistic and Regularization," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics* **34**, pp. 898-911, 2004.

[15] fastkmeans Matlab code, C. Elkan, Univ. of Calif., San Diego.

[16] S. Rein, F. Fitzek, M. Reisslein, "Voice Quality Evaluation in Wirelesss Packet Communication Systems: A Tutorial and Performance Results for ROHC," *IEEE Wireless Communications*, pp. 60-67, February 2005; and "Voice Quality Evaluation for Wireless Transmission with ROHC, extended version," *Technical Report acticom-03-002*, Arizona State Univ., Tempe, AZ, 2003.

[17] http://www.neurones.espci.fr/denby/ouisper.php