EIGENTONGUE FEATURE EXTRACTION FOR AN ULTRASOUND-BASED SILENT SPEECH INTERFACE

T.Hueber[1,3], G.Aversano[3], G.Chollet[3], B.Denby[1,2], G.Dreyfus[1], Y.Oussar[1], P.Roussel[1], M.Stone[4]

[1]Laboratoire d'Electronique, Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI-Paristech), 10 rue Vauquelin, 75231 Paris Cedex 05 France ; thomas.hueber@gmail.com
[2]Université Pierre et Marie Curie – Paris VI, B.C. 252, 4 place Jussieu, 75252 Paris Cedex 05, France ; denby@ieee.org
[3]Laboratoire Traitement et Communication de l'Information, Ecole Nationale Supérieure des Télécommunications (ENST-Paristech), 46 rue Barrault, 75634 Paris Cedex 13
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore MD 21201 USA

ABSTRACT

The article compares two approaches to the description of ultrasound vocal tract images for application in a "silent speech interface," one based on tongue contour modeling, and a second, global coding approach in which images are projected onto a feature space of *Eigentongues*. A curvature-based lip profile feature extraction method is also presented. Extracted visual features are input to a neural network which learns the relation between the vocal tract configuration and line spectrum frequencies (LSF) contained in a one-hour speech corpus. An examination of the quality of LSF's derived from the two approaches demonstrates that the eigentongues approach has a more efficient implementation and provides superior results based on a normalized mean squared error criterion.

Index Terms— image processing, speech synthesis, neural network applications, communication systems, silent speech interface

1. INTRODUCTION

There has been significant interest recently in the notion of a "silent speech interface (SSI)" – a portable device used as an alternative to tracheo-oesophageal speech for larynx cancer patients, for situations where silence must be maintained, or for voice communication in noisy environments. Approaches based on electromyography [1], a non-audible murmur microphone [2], and ultrasound and optical imagery ([3], [4]) have appeared in the literature.

We present here results of a visuo-acoustic SSI study based on a one-hour corpus comprising ultrasound and optical imagery of the vocal tract. The use of a corpus of this size – which was motivated by the desire to interface to a concatenative speech synthesizer – has led to the development of robust feature extraction techniques in order to accommodate the wide variety of articulator configurations appearing in the corpus. In particular, an *Eigentongues* approach has been introduced in order to address the problem of ultrasound frames in which the tongue images poorly. Section 2 of the article details data acquisition and ultrasound image preprocessing, while section 3 describes the feature extraction techniques used in the image (ultrasound and optical) and speech signal analyses. Modeling of the link between visual and acoustic features is introduced in section 4, along with experimental results.

2. DATA ACQUISITION AND PREPROCESSING

2.1. Data acquisition

Data were taken using a 30 Hz ultrasound machine and the Vocal Tract Visualization Lab HATS system [5], which maintains acoustic contact between the throat and the ultrasound transducer during speech. A lip profile image is embedded into the ultrasound image, as shown in figure 1.

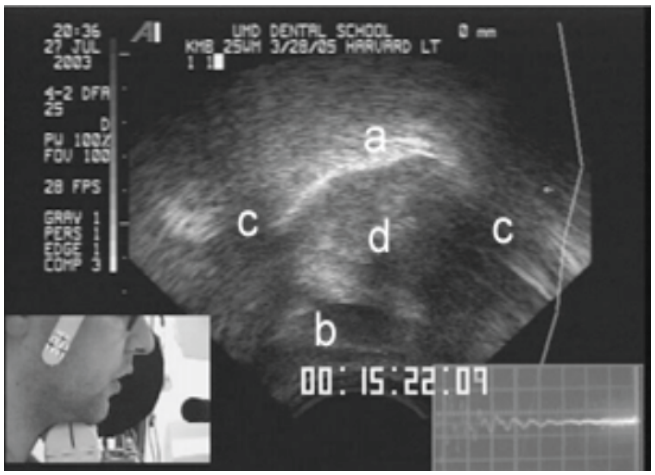

Figure 1. Example of an ultrasound vocal tract image with embedded lip profile : (a) tongue surface ; (b) hyoid bone ; (c) hyoid and mandible acoustic shadows ; (d) muscle, fat and connective tissue within the tongue.

The speech dataset used consists of 720 sentences, organized in 10 lists, from the IEEE/Harvard corpus [6], spoken by a male native American English speaker. The IEEE sentences were chosen because they are constructed to have roughly equal intelligibility across lists and all have approximately the same duration, number of syllables, grammatical structure and intonation. After cleaning the database, the resulting speech was stored as 72473 JPEG frames and 720 WAV audio files sampled at 11025 Hz.

2.2. Ultrasound image preprocessing

In order to select a region of interest, the ultrasound images are first reduced to a 50 (radial angle) by 50 (azimuthal angle) semi-polar grid. To decrease the effects of speckle, the reduced images are filtered using the anisotropic diffusion filter proposed by Yu [7]. This iterative process introduces intra-region smoothing while inhibiting inter-region smoothing [8], *via* a local coefficient of variation [9], so that speckle is removed without destroying important image features. A typical result after these two preprocessing steps is illustrated in figure 2(a).

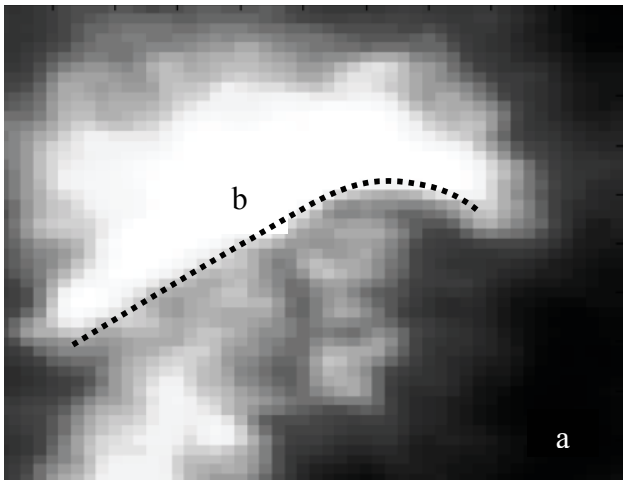

Figure 2. Reduced and filtered ultrasound image (a) and tongue surface contour fit by a 4th order spline (b)

3. FEATURE EXTRACTION

3.1. Ultrasound image feature extraction

3.1.1. Tongue contour extraction

As in [3] and [4], our first approach considers the tongue surface to be the only ultrasound image information relevant to the prediction of speech characteristics. Tongue contour candidate points are defined as maxima of the smoothed vertical intensity gradient. Then, in the present work, a Least Median Square (LMS, [10])-based spline interpolation method, tolerating up to 50% outlier points, is used in order to retain only relevant tongue contour candidates; this is an improvement over the contour extraction method implemented in [3] and [4].

A typical tongue contour is shown in figure 2(b). Due to refraction, however, the tongue surface will be poorly imaged when the tongue surface is at angles nearly parallel to the ultrasound beam, as in the case of the phoneme /i/ for example. The contour extraction described previously fails in such frames – which are found to constitute some 15 % of our database – since the tongue surface is simply no longer visible in them. These "outlier frames" are detected automatically using the area of the convex hull of intensity gradient maxima. Below, we present a more global feature extraction approach which provides a solution to the missing contour problem.

3.1.2. Eigentongue feature extraction

The second approach features the use of Principal Component Analysis (PCA), or Karhunen-Loève expansion, for describing the ultrasound images. The first step is to create a finite set of orthogonal images, which constitutes, up to a certain accuracy, a subspace for the representation of all likely tongue configurations. These images are referred to as *Eigentongues*, a term inspired by the *Eigenface* method of Turk and Pentland [11]. The first three *Eigentongues*, obtained after a PCA on 1000 reduced and filtered ultrasound images, are shown in figure 3.

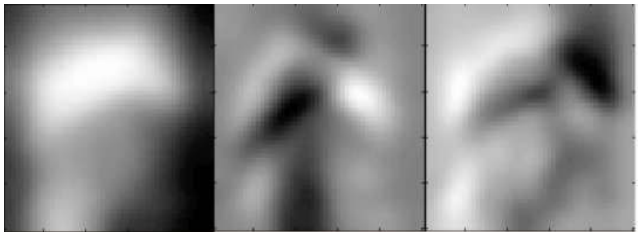

Figure 3. The first three *Eigentongues* (1-3 from left to right)

Once the set of *Eigentongues* has been created, the images of subsequent tongue configurations can be represented quite compactly in terms of their projections onto the set of *Eigentongues*, as shown in figure 4.

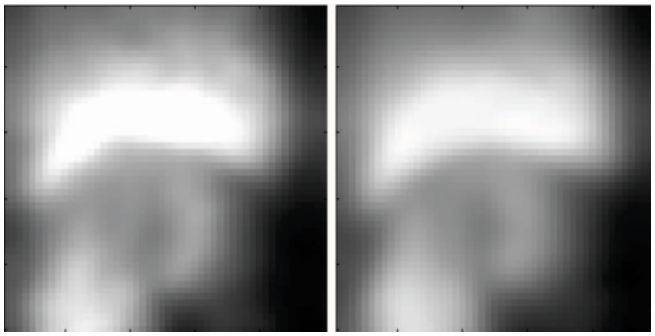

Figure 4. A reduced ultrasound image (left) and its re-synthesis (right) using 20 *Eigentongue* components

The *Eigentongue* components encode the maximum amount of relevant information in the images, mainly tongue position, of course, but also other structures such as the hyoid bone, muscles, etc.

3.2. Optical image feature extraction

The optical image feature extraction consists of a description of the lip profile. We propose an algorithm based on the observation of Attneave that information along a visual contour is concentrated in regions of high curvature, rather than distributed uniformly [12]. The lip edge profile is easily extracted using the Sobel method. The curvature of this two-dimensional curve is then computed using the Turning Angle introduced by Feldman [13]. Upper/lower lip and commissure positions coincide with extrema of the curvature, as shown as figure 5, while the values of the curvature at these points give local lip shape information.

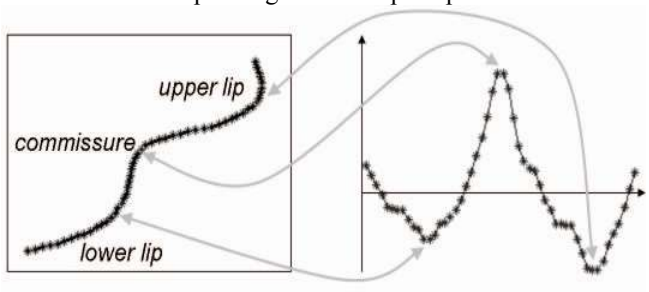

Figure 5. Lip profile description using curvature computation (left: lip contour; right: curvature of lip contour)

3.3. Speech signal description

For each 33 ms audio frame (dictated by the 30 Hz ultrasound rate), twelve LSF's are calculated using a pre-accentuation filter, linear predictive coding and a Hann window with a half-frame overlap. The robustness of LSF coefficients is known to assure the stability of the LPC filter [14]. A voiced/unvoiced flag and fundamental frequency (for voiced frames) are also computed, using a simple autocorrelation-based method. These last two features are not used in the visuo-acoustic modeling which follows, but allow a qualitative, audible comparison of our different results, if desired, *via* LPC synthesis using the predicted autoregressive filter coefficients and a realistic excitation function.

4. VISUO-ACOUSTIC MODELING

Our first feature extraction method, described in sections 3.2 and 3.1.1, provides 15 features per frame, including 9 for the lips (position and curvature of upper/lower lips and commissure) and 6 for the tongue (4^{th} order spline coefficients and interval of definition). The second, *Eigentongue* method gives 29 features per frame, the first 20 *Eigentongue* components plus lip features. A multilayer perceptron (MLP) is used to perform the mapping between these input visual features and the 12 LSF's [15]. A separate network is used for each LSF in order to limit the number of adjustable parameters in the model. A total of 71502 frames are used for training, with an independent set of 971 frames for model selection and validation. We now compare the LSF prediction obtained from the two methods. Because each LSF is defined upon its own interval, we introduce a normalized measure of the quality of the prediction α, along with an estimate of its standard deviation ε [16]:

$$\alpha = \frac{100}{\sqrt{N}} \cdot \frac{\sqrt{\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2}}{|y_{\max} - y_{\min}|} \qquad \varepsilon = \alpha \sqrt{\frac{1}{2N}}$$

where N is the number of examples in the validation database, y are the true LSF's, and $\tilde{y}$ the predicted LSF's.

4.1. Comparing Contour and *Eigentongue* approaches

For the tongue contour method, in order to obtain reasonable training results, the "outlier frames" for which the automatic contour extraction algorithm (described in section 3.1.1) failed were removed from the training set. As the *Eigentongue* feature extraction approach does not restrict relevant information to a specific structure, no outlier is generated when that structure is not imaged, and thus all frames may be used in the visuo-acoustic modeling with this method. Columns 1 and 2 of Table 1 compare the results of the two approaches.

LSF number	Tongue Contour	*Eigentongue*	*Eigentongue* + history
	Parameter α (% of total dynamic range)		
1	18.7 ± 0.4	16.9 ± 0.4	16.4 ± 0.4
2	16.1 ± 0.4	14.4 ± 0.3	13.7 ± 0.3
3	14.3 ± 0.3	12.4 ± 0.3	12.3 ± 0.3
4	13.1 ± 0.3	11.8 ± 0.3	10.8 ± 0.2
5	14.2 ± 0.3	11.5 ± 0.3	11.9 ± 0.3
6	13.1 ± 0.3	11.8 ± 0.3	10.6 ± 0.2
7	15.7 ± 0.4	13.7 ± 0.3	12.6 ± 0.3
8	13.1 ± 0.3	11.8 ± 0.3	12.1 ± 0.3
9	14.6 ± 0.3	12.8 ± 0.3	12.4 ± 0.3
10	12.9 ± 0.3	11.2 ± 0.2	11.2 ± 0.2
11	14.5 ± 0.3	13.7 ± 0.3	11.4 ± 0.2
12	16.3 ± 0.4	14.5 ± 0.3	14.4 ± 0.3

Table 1. Comparison of tongue contour based modeling and *Eigentongue* based modeling. Quoted errors, ε, are estimates of the standard deviation of α using a Gaussian assumption

The table shows that LSF's 4, 6, 8 and 10 are the best predicted by tongue contour and lip profile features, and that using *Eigentongues* provides an improvement in overall prediction quality which is small, but statistically

significant. The filtering step described in section 2.2 is in fact not essential for the *Eigentongue* feature extraction, as image regions of high intensity variability will be associated with the higher order *Eigentongues*, which are not used. Similar results are obtained using *Eigentongues* obtained from unfiltered images.

4.2. Introducing 'history' into the input variables

The use of *Eigentongues* allows all of the video frames to participate in the training, which is not the case for the contour method due to the missing frames. We can then in a simple way take account of the intrinsically dynamic nature of speech production in our visuo-acoustic modeling by providing the training algorithm, at frame *n*, with the *Eigentongue* and lip variables of frames *n*-1 and *n*-2, as well. An additional small improvement in the prediction of LSF's 2, 4, 6, 7 and 11 is seen, as compared to the static modeling.

5. CONCLUSION AND PERSPECTIVES

A new turning-angle algorithm for the description of lip profiles has been introduced, which, because curvature-based, should hopefully make the method robust against the variability of lip shapes between speakers. Two methods for feature extraction from ultrasound images have been presented and compared. The visuo-acoustic modeling with *Eigentongues* gives better results than those obtained using tongue contours as input. The *Eigentongue* method is easier to implement, appears to take more information into account, and is not prone to failures due to instrumental effects, thus allowing the dynamic nature of speech to be taken into account in a natural way. It could be interesting, however, in future work, to combine the two approaches in the context of active appearance models [17]. The model we propose is at present able to predict an acoustical description of speech with errors ranging from 11% to 16%. Whether this performance is adequate for application in an SSI will only become apparent once a concatenative speech synthesis model using our predicted quantities as inputs has been experimented. The elaboration of such a test, as well as the use of alternative dynamic process modeling techniques (Hidden Markov Models, Time Delay Neural Networks [18]) are currently underway.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge useful discussions with Isabelle Bloch. This work was supported in part by CNFM (Comité National de Formation en Microélectronique).

7. REFERENCES

[1] C. Jorgensen, D. D. Lee, and 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, 2003.

[2] Y. Nakajima, P. Heracleous, H. Saruwatari and K. Shikano, "A Tissue-conductive Acoustic Sensor Applied in Speech Recognition for Privacy," *Smart Objects & Ambient Intelligences Oc-EUSAI 2005*, pp. 93-98, 2005.

[3] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone, "Prospects for a Silent Speech Interface Using Ultrasound Imaging," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.

[4] B. Denby and M. Stone, "Speech Synthesis from Real Time Ultrasound Images of the Tongue," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.

[5] M. Stone, "A Guide to Analysing Tongue Motion from Ultrasound Images," *Clinical Linguistics and Phonetics*, pp. 359-366, 2003.

[6] IEEE, "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225-246, 1969.

[7] Y. Yu and S. T. Acton, "Speckle Reducing Anisotropic Diffusion," *IEEE Transactions on Image Processing*, vol. 11, pp. 1260-1270, 2002.

[8] P. Perona and J. Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 629-639, 1990.

[9] J. Lee, "Digital Image Enhancement and Noise Filtering by Use of Local Statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, pp. 165-168, 1980.

[10] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, New York, USA, John Wiley & Sons, Inc., 1987.

[11] M. A. Turk and A. P. Pentland, "Face Recognition Using Eigenfaces," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586-591, 1991.

[12] F. Attneave, "Some Informational Aspects of Visual Perception," *Psych. Review*, vol. 61, pp. 183-193, 1954.

[13] J. Feldman and M. Singh, "Information Along Contours and Object Boundaries," *Psych. Review*, vol. 112, pp. 243-252, 2005.

[14] G. Kang and L. Fransen, "Application of Line-Spectrum Pairs to Low-Bit-Rate Speech Encoders," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tampa, USA, 1985.

[15] G. Dreyfus, *Neural Networks: Methodology and Applications*, Springer, New York, 2005.

[16] H. Cramér, *Mathematical Methods of Statistics,* Princeton University Press, Princeton, 1999.

[17] M. B. Stegmann, R. Fisker, B. K. Ersbøll, H. H. Thodberg, L. Hyldstrup, "Active Appearance Models: Theory and Cases," *Proc. 9th Danish Conference on Pattern Recognition and Image Analysis*, vol. 1, pp. 49-57, AUC Press, 2000.

[18] T. Hanazawa, A. Waibel, G. Hinton, K. Shikano, and K. Lang, "Phoneme Recognition Using Time Delay Neural Networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 328-339, 1989.