Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips

Thomas Hueber [1,3], *Gérard Chollet* [3], *Bruce Denby* [2,1], *Gérard Dreyfus* [1], *Maureen Stone* [4]

[1]Laboratoire d'Electronique, ESPCI Paristech, 10 rue Vauquelin, 75231 Paris Cedex 05 France
[2]Université Pierre et Marie Curie - Paris VI, 4 place Jussieu, 75252 Paris Cedex 05 France
[3]Laboratoire Traitement et Communication de l'Information, Telecom Paristech, 46 rue Barrault, 75634 Paris Cedex 13 France
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore, MD 21201 USA

hueber@ieee.org, gerard.chollet@tsi.enst.fr, denby@ieee.org, gerard.dreyfus@espci.fr, mstone@umaryland.edu

Abstract

This article presents a framework for a phonetic vocoder driven by ultrasound and optical images of the tongue and lips for a "silent speech interface" application. The system is built around an HMM-based visual phone recognition step which provides target phonetic sequences from a continuous visual observation stream. The phonetic target constrains the search for the optimal sequence of diphones that maximizes similarity to the input test data in visual space subject to a unit concatenation cost in the acoustic domain. The final speech waveform is generated using "Harmonic plus Noise Model" synthesis techniques. Experimental results are based on a one-hour continuous speech audiovisual database comprising ultrasound images of the tongue and both frontal and lateral view of the speaker's lips.

Index Terms: silent speech, corpus-based speech synthesis, visual speech recognition

1. Introduction

The objective of a "Silent Speech Interface" is to permit voice communication without the vocalisation of sound. Such a system primarily targets applications in which silence must be maintained, but could also be used to enable voice communication in situations where standard speech is masked by background noise. Since no glottal activity is required, it could furthermore have application as an alternative to tracheo-oesophageal and electrolaryngeal speech for laryngectomized patients. In the literature, silent communication has usually been envisioned as a speech recognition task driven by observation of the voice organ. The input articulator activity may be derived from EMG/EPG signals, as in [1], or, as in our case, from ultrasound and optical images of the vocal tract.

In [2] and [3], we addressed the problem of continuous-speech phone recognition from ultrasound and optical video sequences of the vocal tract. Here, we propose to use this visual phone recognition step (VSR) as the basis of a phonetic vocoder driven by video-only data. Our approach does not use a specific vocal tract model as in articulatory synthesis, but rather is based on the building of an audiovisual dictionary in which each visual unit has an equivalent in the acoustic domain. Given a test sequence of visual features and the phonetic target predicted by the VSR, a unit selection algorithm searches in this audiovisual dictionary the optimal sequence of units that best matches the input test data. The proposed unit selection algorithm is an adaptation of the standard path search algorithm used in corpus-based speech synthesis. The quality of the match is defined optimally as a compromise between a target cost evaluated in the visual space and a concatenation cost evaluated in the acoustic domain. The output speech waveform is generated by concatenating a "Harmonic plus Noise Model" (HNM) representation of acoustic segments for all selected units. An overview of the recognition/synthesis system is given in figure 1.

The system is evaluated on a 61 minute audiovisual database of ultrasound and optical sequences of the tongue and lips, recorded in synchrony with the uttered speech signal. Text material was chosen with corpus-based synthesis specifically in mind.

Section 2 of the article summarizes database content and acquisition, feature extraction procedures, and the visual phone recognition step (further details on these system blocks are given in [3]). The unit selection algorithm and speech waveform generation techniques used in the corpus-based synthesis, which are the main focus of this article, are detailed in section 3, along with preliminary experimental results of synthesis driven by video-only data.

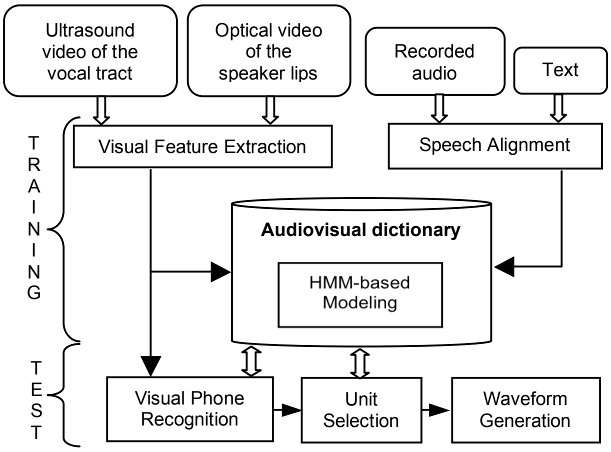

Figure 1: *Framework for a phonetic vocoder based on visual observation of the tongue and lips.*

2. Visual Phone Recognition Step

2.1. Database acquisition and phonetic content

The acquisition system fixes the speaker's head and supports the ultrasound transducer under the chin without disturbing articulator movement [4]. The protocol described in [2] was modified to include both lateral and frontal view of the speaker's lips along with the ultrasound tongue images and acoustic speech signal. The streams are mixed at a video frame rate of 30 Hz. A typical frame is shown in figure 2.

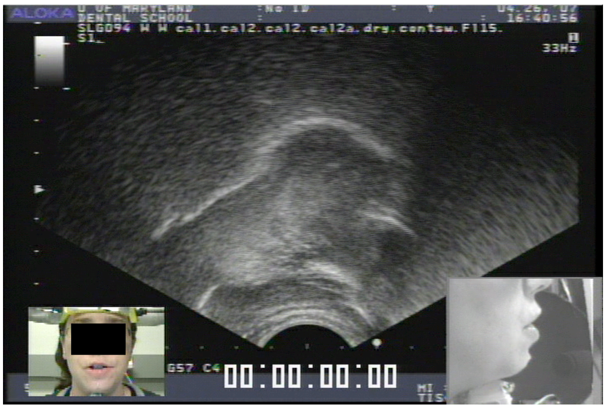

Figure 2: *Example ultrasound vocal tract image showing frontal and lateral lip views*

Because the recorded multimodal speech signal is used both for phone-based VSR and as the basis of a diphone-based concatenative synthesizer, the text of the database must be phonetically balanced and have good diphone coverage. The CMU-Arctic corpus text [5] was chosen for our acquisitions. This base consists of 1132 sentences divided into two phonetically balanced sets, A and B, of 593 and 539 items respectively. With a phoneme set of 41 elements (39 phonemes plus schwa and pause), the diphone coverage of sets A and B in the corpus is 78 % and 75.4 % respectively.

A native speaker of American English read sentences from sets A and B in a single session lasting over 2 hours. Speaker fatigue limited acquisitions to the first 1020 of the 1132 Arctic phrases (100 % of set A and 80 % of set B). Multiple sessions are not done at present in order to avoid compensating for imprecisions in the re-alignment of the transducer relative to the vocal tract. After cleanup, the resulting 61 minutes of speech was stored as 109553 bitmap frames and 1020 WAV audio files sampled at 16000 Hz.

2.2. Phonetic alignment of the speech waveform

The acoustic wave of each recorded sentence is parameterized by 12 Mel-frequency cepstral coefficients (MFCC) with their energies and first and second derivatives. The phonetic labeling is performed by an HMM-based forced alignment procedure with an initial set of 40 HMM acoustic models, trained on the transcribed multi-speaker DARPA TIMIT speech database [6]. These 5-state (with one non-emitting initial state, and one non-emitting terminating state), 16 mixture, left-to-right HMM models are used to segment the corpus audio stream. The HMM training and recognition procedure are done using the HTK front-end [7].

To facilitate the subsequent diphone speech synthesis, a segmentation of the database into diphones is deduced from the phone labeling by searching for spectral stability points at the boundaries of all phones. With 33637 phonemes labeled, the diphone coverage of our audiovisual speech database is found to be 79.4 % (1271 diphones out of a total of 1599).

2.3. Tongue and lip video feature extraction

Tongue and the lip regions of interest are resized to 64x64 pixels via cubic interpolation and ultrasound images are filtered using an anisotropic diffusion filter [8]. Then the PCA-based EigenTongues/EigenLips decomposition [9] is used to encode the tongue and lips by considering their positions as a linear combination of standard configurations. The number of useful EigenTongues and EigenLips coefficients to keep is fixed empirically at 20 and 15 respectively. Features are finally resampled at 100 Hz using linear interpolation before being concatenated with first and second derivatives into a single vector.

2.4. Recognition protocol and performance

Observed sequences of each phonetic class are modeled by a left-to-right, 5-state (3 emitting states, 32 gaussians per state), continuous monophone HMMs. No statistical language model is used at this stage so as to allow evaluation of the quality of the HMM-based modeling alone. The database is divided into 34 lists of 30 sentences. In the performance estimation, a jackknife strategy [10] uses each list once for validation while the other 33 make up the training set.

The recognizer performance P_{VSR} is defined as

$$P_{VSR} = 100 \cdot \frac{N - D - S - I}{N} \tag{1}$$

where N is the total number of phones in the test set, S the number of substitution errors, D deletion errors, and I insertion errors. To establish a performance target for the visual recognizer, a standard acoustic-based phone recognizer is evaluated on the same database. This uses 40 context-independent, left-to-right, 5-state, 16-mixture, continuous monophone HMMs estimated on each training pass.

Table 1 compares performances of the visual-based and acoustic-based phone recognizers. VSR performance is already almost 80 % of that obtained using ASR, indicating that initial synthesis experiments are indeed justified. A full discussion of the visual phone recognition step is given in [3].

	ASR	VSR = Tongue + Lips	
		Lateral	Frontal
$P \pm \Delta$	83.9 ± 0.7 %	56 ± 0.9 %	**60 ± 0.9 %**

Table 1. *Visual and acoustic based phone recognizer performance P for a 95% confidence interval Δ*

3. Corpus-Based Synthesis

3.1. Unit selection

The visual speech recognizer is able to identify a discrete sequence of phones in a continuous stream of visual features. The recorded database, automatically labeled at the phonetic level (section 2.2), can in turn be considered as an audiovisual dictionary of speech units in which each visual item has an equivalent in the acoustic domain. In our proposed phonetic vocoder, the visual phone recognizer drives corpus-based synthesis assisted by a unit selection procedure. Starting from the predicted phonetic target, the algorithm searches the optimal sequence of diphones that maximize similarity to input test data in visual space while limiting unit concatenation cost in the acoustic domain. This algorithm is based on the standard path search algorithm used in concatenative speech synthesis described in [11]. The overall scheme is illustrated in figure 3.

Assuming a test sequence of visual features $v = v_1 \ldots v_N$ where N is the length of sequence, and $\tau = \tau_{1 \ldots} \tau_T$ the temporal segmentation of v given by the visual phone recognizer, the sequence t_τ of T target units is defined by:

$$t_\tau = [t_{\tau_1}],\ldots,[t_{\tau_T}] = [v_{\tau_1},\ldots,v_{\tau_2}],\ldots,[v_{\tau_{T-1}},\ldots,v_{\tau_T}] \quad (2)$$

The unit selection algorithm finds, among all appropriate units, the optimal sequence $\{u_k\}$ that best matches the target t_τ. The quality of the match is determined by two costs, C^t and C^c.

The target cost C^t expresses the visual similarity between target units and the units selected in the dictionary and is given by:

$$C^t(u_k, t_{\tau_i}) = DTW(u_k, t_{\tau_i}) \quad (3)$$

where $DTW(a,b)$ is the cumulative distance obtained after a dynamic time warping between the two sequences of visual feature vectors. This non-linear alignment procedure takes naturally into account temporal stretching and compression of the motion of the articulators.

The concatenation cost C^c estimates the spectral discontinuity introduced by the concatenation of two units u_{LEFT} and u_{RIGHT} and is given by:

$$C^c(u_{LEFT}, u_{RIGHT}) = D\left(MFCC(u_{LEFT_{END}}), MFCC(u_{RIGHT_1})\right) \quad (4)$$

where D is the Euclidean distance and $MFCC(u_l)$ are MFCC coefficients of the unit u at frame l.

Because the audiovisual dictionary can be considered as a fully connected state transition network, the search for the least costly path that best matches the test sequence can be determined by a Viterbi algorithm [12]. In this network, each state is occupied by a unit. State occupancy is estimated using the visual-based target cost function and transition between states is evaluated by the acoustic-based concatenation cost.

3.2. HNM-based speech waveform generation

After the selection stage, speech can be synthesized by concatenating acoustic components of selected diphones. However, because no prosodic information such as pitch, energy and duration, is used during the unit selection stage, pitch and time-scale adaptations are necessary. Acoustic modifications are achieved using a "Harmonic Plus Noise" representation of the speech signal [13]. In the HNM framework, the spectrum of a speech frame $s(t)$ is described as the sum of a harmonic part $H(t)$ and a noise part $B(t)$:

$$s(t) = H(t) + B(t) = \left[\sum_{k=1}^{N} A_k \cos\left(2\pi k f_0 t\right) + \varphi_k\right] + \left[N_{gauss} * F(t)\right] \quad (5)$$

where N is the number of harmonics included in $H(t)$, f_0 is the estimated fundamental frequency, N_{gauss} a gaussian noise frame and $F(t)$ an autoregressive filter. Our implementation employs 12 harmonic components along with a 16th-order auto-regressive model for the noise part.

HNM is a pitch-synchronous scheme that is flexible enough to implement good-quality prosodic modifications. In our case, acoustic modifications consist of phone duration adaptation, and pitch and spectral smoothing. Phone durations are adapted according to the temporal segmentation provided by the HMM-based phone recognition step described in section 2. Because no information about the global evolution of pitch is directly available in a silent speech application (absence of glottal activity), a strong smoothing of the fundamental frequency over the sentence is applied. Such a basic treatment helps limit non-realistic prosodic variations but (empirically) can degrade voice naturalness. As a final step, the HNM parameters are smoothed near diphone boundaries using linear interpolation.

The example chosen for figure 3 illustrates the interplay between the two cost functions. The diphone [w-ih] is selected correctly for its similarity to the test sequence. However, the next diphone [ih-ah] does not match well with the input sequence (as at the end of phone [ah]); the selection of this unit is mainly due to its acoustic continuity with the previous unit. We note that in the present algorithm, the target and concatenation cost are weighted manually.

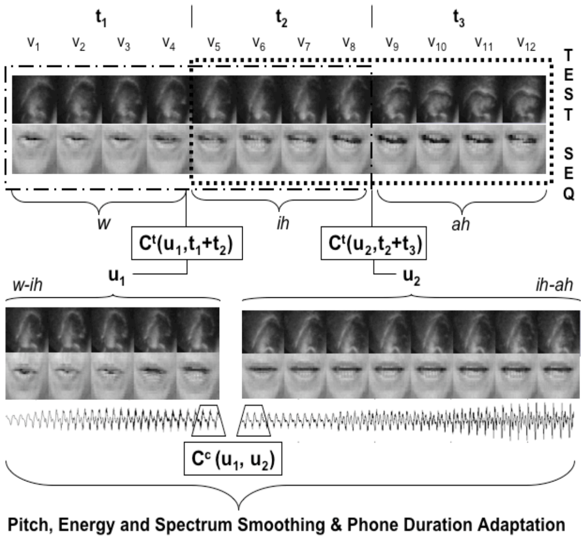

Figure 3: *Corpus-based synthesis procedure (T=3)*

3.3. Experimental results

The quality of the synthesized waveform obviously depends strongly on the performance of the visual phone recognizer. In our current framework, the unit selection synthesis is driven exclusively by the predicted phonetic sequence, and thus an error during the recognition stage will necessarily corrupt the synthesis. With visual phone recognizer accuracy of only 60%, consistently intelligible synthesis is not yet possible.

A first empirical evaluation of our "silent vocoder" is presented in figure 4. Example 1 represents a 'typical' performance of the system, with 69 % of the phones correctly identified, while the phrase of example 2 has 95 % of phones correctly matched. Two distinct types of errors are apparent (see also [3]): first, phones with similar articulatory gestures, such as {[p],[b],[m]} are sometimes confused; secondly, very short phones such as {[t],[b],[n]} can be missed due to the 30 Hz acquisition rate. The multimedia file provided for the second example illustrates the ability of the synthesis technique to produce an intelligible speech signal with "acceptable" prosody when the predicted phonetic target is correct. There are still difficulties identifying short pauses or within-sentence silences (anticipation phenomena), and better results are obtained on short sentences with no more than one or two prosodic groups. Clearly, a more detailed study of the impact of different types of error on synthesis quality will be necessary. Thus, although our system is still not fully functional, this approach for a segmental speech coder driven only by visual observation seems promising.

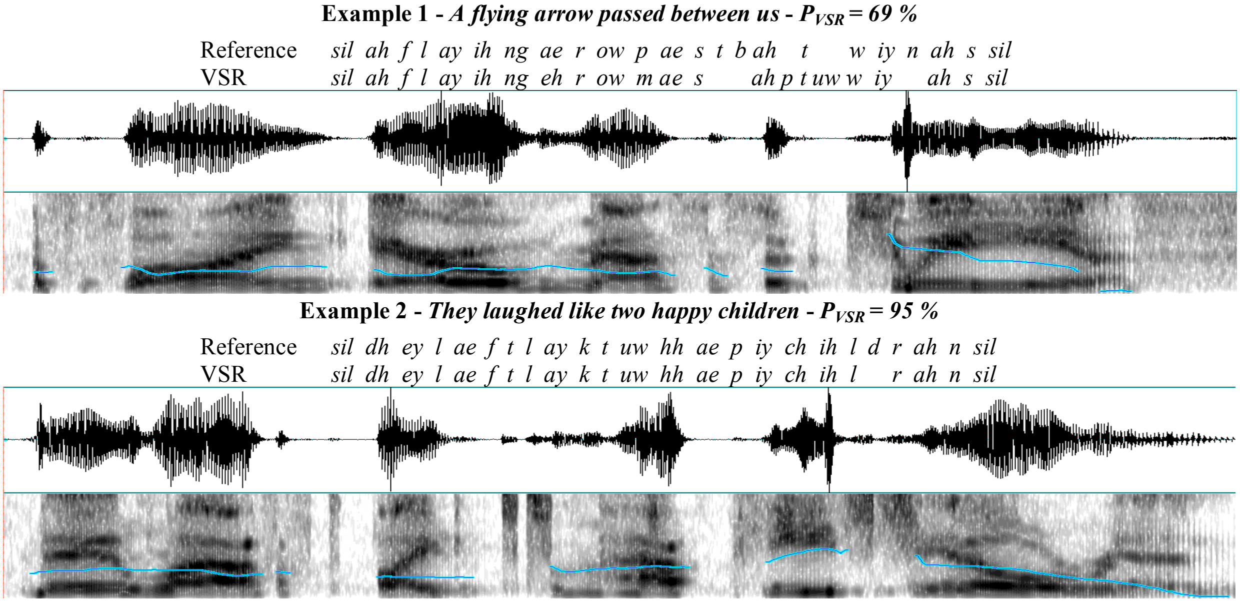

Figure 4: *Phone recognition and associated corpus-based synthesis from video-only data (fundamental frequency in light blue)*

Multimedia files submitted

"exampleX_synth.wav" with X=1,2 : Synthesis from video-only speech data

"exampleX_orig.wav" with X=1,2 : Original sentence (target)

4. Conclusions and Perspectives

The proposed segmental speech coder driven by video-only data combines an HMM-based visual phone recognition stage with an audiovisual unit selection algorithm and robust HNM-based synthesis techniques. To date, synthesis quality depends only on the performance of the visual phone recognizer, currently at 60 %. To improve the recognition stage, several solutions are envisioned. The use of a statistical language model or phonotactic linguistic constraints in the HMM decoding stage will be investigated. A new, higher rate acquisition system is also under development in order to reduce the number of phone deletion errors. As our modeling technique does not presently take into account possible asynchronies between articulators, the use of multistream HMMs [14] could prove useful. The unit selection algorithm furthermore is currently driven only by the output of the phone-based recognizer; it might be more fruitful to consider a combination of HMM-based stochastic modeling and data-driven techniques. A deeper dictionary search, also including longer units, such as polyphones, could capture more contextual effects and improve general performance. Finally, the system should be evaluated on more realistic test databases containing either whispered or totally silent speech. Such data will be very useful to learn about the particularities of tongue and lip movement in silent speech.

5. Acknowledgements

This work is supported by the French Department of Defense (DGA) and the French National Research Agency, under contract number ANR-06-BLAN-0166.

6. References

[1] Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., "Session independent non-audible speech recognition using surface electromyography," IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 331-336, 2005.

[2] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips," Interspeech, pp. 658-661, Antwerp, Belgium, 2007.

[3] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Visual Phone Recognition for an Ultrasound-Based Silent Speech Interface," *submitted to these proceedings*.

[4] Stone, M., and Davis, E., "A Head and Transducer Support System for Making Ultrasound Images of Tongue/Jaw Movement," Journal of the Acoustical Society of America, vol. 98 (6), pp. 3107-3112, 1995.

[5] Black, A. W., Lenzo, K., "Building voices in the Festival speech synthesis system," *http://festvox.org/bsv*, 2000.

[6] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NTIS order number PB91-100354, 1993.

[7] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., The HTK Book, September 2005, http://htk.eng.cam.ac.uk/.

[8] Y. Yu and S. T. Acton, "Speckle Reducing Anisotropic Diffusion," IEEE Trans. on Image Proc., vol. 11, pp. 1260-1270, 2002.

[9] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," ICASSP, Honolulu, pp. I1245-I1248, 2007.

[10] Efron, B., "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," Biometrika, vol. 68, pp. 589-599, 1981.

[11] Hunt, A. J., Black, A. W., "Unit selection in a concatenative speech synthesis system using a large speech database," IEEE ICASSP, pp. 373–376, Atlanta, 1996.

[12] Forney, G. D., The Viterbi algorithm. Proceedings of the IEEE 61(3), pp. 268-278, 1973.

[13] Stylianou, Y., Dutoit, T., Schroeter, J., "Diphone Concatenation using a Harmonic plus Noise Model of Speech," Eurospeech, pp. 613-616, Rhodes, Greece, 1997.

[14] Gravier, G., Potamianos, G., and Neti, C., "Asynchrony modeling for audio-visual speech recognition," In Proceedings of the Second international Conference on Human Language Technology Research, San Diego, California, 2002.