Phone Recognition from Ultrasound and Optical Video Sequences for a Silent Speech Interface

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

Latest results on continuous speech phone recognition from video observations of the tongue and lips are described in the context of an ultrasound-based silent speech interface. The study is based on a new 61-minute audiovisual database containing ultrasound sequences of the tongue as well as both frontal and lateral view of the speaker's lips. Phonetically balanced and exhibiting good diphone coverage, this database is designed both for recognition and corpus-based synthesis purposes. Acoustic waveforms are phonetically labeled, and visual sequences coded using PCA-based robust feature extraction techniques. Visual and acoustic observations of each phonetic class are modeled by continuous HMMs, allowing the performance of the visual phone recognizer to be compared to a traditional acoustic-based phone recognition experiment. The phone recognition confusion matrix is also discussed in detail.

Index Terms: silent speech interface, visual speech recognition

1. Introduction

In recent years, the design of devices allowing silent verbal communication has emerged as a new field in speech processing research. Such "Silent Speech Interfaces" (SSI) could be useful for voice communication in situations where silence must be maintained, or, conversely, in very noisy environments. An SSI might also be an alternative to tracheo-oesophageal or electrolaryngeal speech for laryngeal cancer patients. To build an SSI, voice organ activity could be derived from EMG/EPG signals, as in [1], or, if whispered speech can be tolerated, using a "non-audible murmur microphone" (NAM [2]). In our work, an ultrasound transducer below the chin and a standard video camera (which would be integrated and miniaturized in a final application) are used to directly image the tongue and lips, respectively [3].

In [4], we addressed the problem of continuous-speech phone recognition from ultrasound and optical sequences of the vocal tract as a first step toward corpus-based synthesis. In that work, a visual speech recognizer (VSR) was evaluated on a 43 minute database containing ultrasound tongue images and the lips in profile. The goal of the present article is to evaluate the robustness of our VSR on a larger database with a different speaker. To that end, a new audiovisual database containing 61 minutes of ultrasound and optical sequences of the tongue and lips was recorded together with the uttered speech signal. Corpus text was chosen so that the recorded database would be appropriate for later corpus-based synthesis. The acquisition system was also modified to record both frontal and lateral lip views.

A schematic of the recognition/synthesis system is shown in figure 1. The visual phone recognizer predicts a target phonetic sequence from a continuous stream of visual features used to constrain a unit selection algorithm. This algorithm searches an audiovisual dictionary for the sequence of units which best matches input test data. This article focuses on the visual phone recognizer; the unit selection algorithm, which is an adaptation of the standard path search algorithm used in corpus-based speech synthesis, is described elsewhere [5].

Section 2 describes the acquisition of the new database, details its content and presents the visual feature extraction process. Section 3 details the implementation of the visual speech recognizer and evaluates its performance; a comparison between VSR using frontal and lateral lip images is also presented.

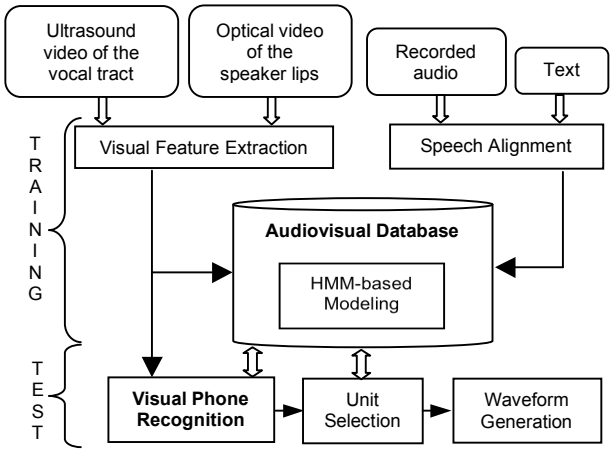

Figure 1: *Framework for corpus-based synthesis driven by visual observation of the tongue and lips.*

2. Building the Audiovisual Speech Database

2.1. Data acquisition protocol and evaluation

Data is recorded using the Vocal Tract Visualization Lab HATS system [6], which maintains the speaker's head immobile and supports the ultrasound transducer under the chin without disturbing speech. The acquisition setup was modified to include two cameras to provide synchronized lateral and frontal view of the speaker's lips together with the ultrasound images of the voice organ and the uttered speech acoustic signal, see figure 2. These three streams are mixed using an analog device, which unfortunately limits the frame rate of the acquisition chain to 30 Hz.

Figure 2: *Example of an ultrasound vocal tract image with embedded lip frontal and lateral view*

Because the recorded multimodal speech signal will serve both for phone-based visual speech recognition and as the basis of a diphone-based concatenative synthesizer, the textual material of the new database must be phonetically balanced and have good diphone coverage. For these two reasons, the CMU-Arctic corpus text [7], which is the basis of the Festvox Text-to-Speech system, was used for our new database. The Arctic database contains 1132 sentences divided into two sets (A and B) containing respectively 593 and 539 items. Both sets are in phonetically balanced American English. Furthermore, with a phoneme set of 41 elements (39 phonemes plus schwa and pause), diphone coverage in sets A and B is 78 % and 75.4 % respectively.

During acquisition, the speaker was instructed to read all sentences of sets A and B as neutrally as possible. Data is recorded in one session during which the speaker remains fixed in the HATS system. Because no re-calibration techniques are employed in our current system, recording data in multiple sessions is not feasible. Since ultrasound imaging of the tongue and its connective tissues (muscle, fat) is very sensitive to modifications of the transducer position, head movement within a session is monitored using palatal traces obtained from 10 cc water deglutitions executed during brief pauses every 90 sentences. During swallowing, the tongue contacts the roof of the mouth, and the ultrasound beam traverses soft tissue until it is reflected by the palate bone [8]. Palatal traces from 4 widely separated deglutitions are shown in figure 3. The proximity of these traces insures that the speaker's head remained stable during the acquisition.

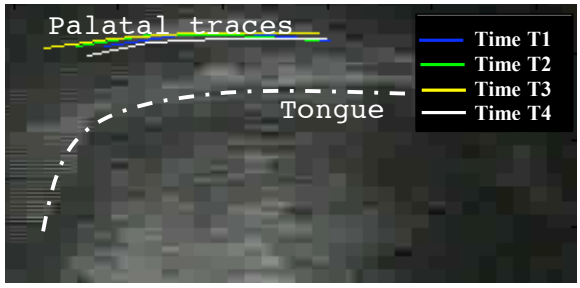

Figure 3: *Superposition of palatal traces extracted from 4 deglutitions recorded periodically during data acquisition.*

The full Arctic A set was acquired, but speaker fatigue, after more than 2 hours in HATS, allowed only 80 % of the B set to be recorded; the total number of sentences was thus 1020 rather than the expected 1132. After cleanup of the database, the resulting 61 minutes of speech was stored as 109553 Bitmap frames and 1020 WAV audio files sampled at 16000 Hz. The new database is thus 30 % larger than that used in our previous study.

2.2. Phonetic alignment of the speech waveform

The acoustic signal of each recorded sentence was first parameterized using 12 Mel-frequency cepstral coefficients, along with their energies and first and second derivatives. The phonetic forced-alignment procedure is a simplified recognition task in which the phonetic sequence is already known. This recognition task is achieved using an initial set of 40 HMM acoustic models trained on the transcribed multi-speaker DARPA TIMIT speech database [9]. These 5-state (with one non-emitting initial state, and one non-emitting terminating state), 16 mixture, left-to-right HMM models are refined on and then used to segment the audio stream of the corpus. All HMM work in our study was done using the HTK front-end [10]. With 33637 phonemes labeled, the actual diphone coverage obtained for our audiovisual speech database was 79.4 % (1271 different diphones of a possible 1599).

2.3. Visual feature extraction

Regions of interest for the tongue and the lips are first resized to 64x64 pixel images using cubic interpolation. In order to decrease the effects of speckle, each ultrasound frame is filtered using an anisotropic diffusion filter [11]. Then, the PCA-based "EigenTongues" decomposition described in [12] is used to encode each frame. An adaptation of the "EigenFaces" method [13], this technique projects each ultrasound image of the vocal tract into the representative space of "EigenTongues", which can be interpreted as the space of the "standard vocal tract configurations". A similar approach is used to code frontal and lateral images of the lips. Figure 4 illustrates how each ultrasound and optical image is coded by its coordinates β_T, β_F, β_L, in the "EigenTongues/EigenLips" space. The indices *n,m,p* which quantify the number of projections onto the set of EigenTongues/EigenLips used for coding are obtained empirically by evaluating the quality of the image reconstructed from its first few components. Typical values of the triplet *(n,m,p)* used for this database are *(20,15,15)*. Finally, visual feature sequences are oversampled from 30 Hz to 100 Hz using linear interpolation. The EigenTongues/EigenLips coefficients, with their first and

second derivative, are concatenated into the same "visual feature vector", in a *feature fusion* strategy.

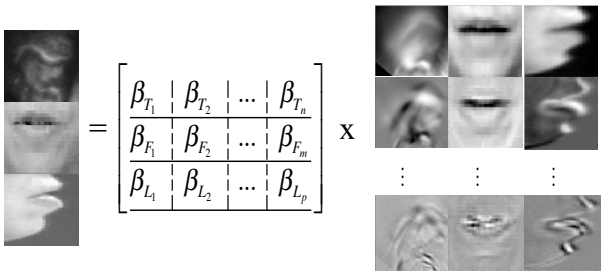

Figure 4: *Encoding ultrasound and optical images of the tongue and lips using the EigenTongue/EigenLips decomposition.*

3. Visual Phone Recognition

3.1. Protocol

The observed visual sequences of each phonetic class are modeled by a left-to-right, 5-state (3 emitting states), continuous HMM (monophone only, due to our dataset size). Models parameters are estimated and refined using incremental embedded training during which the number of Gaussians per state is increased up to 32. As our experiment is intended to show the quality of the HMM-based modeling, neither a statistical language model nor phonotactic constraints are used in this study.

The 1020 sentences of the database are divided into 34 lists of 30 sentences. During performance estimation, each list is used once as the test set while the other 33 lists compose the training set, using a jackknife strategy [14]. The recognizer performance P_{VSR} is defined as

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

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. Although frontal and lateral views of speaker's lips are available in the newly recorded database, the two streams are not used together in the visual phone recognition process, to simulate the conditions of a simple, wearable prototype. A comparative study of visual phone recognition using ultrasound and frontal or lateral lips is however made.

A traditional, acoustic-based phone recognizer is also evaluated on the same database using the HMM acoustic models estimated for the phonetic alignment of the audio-visual database in section 2.2. The performance of this acoustic-based phone recognizer is considered as a 'target' for VSR on this database.

3.2. Results and Interpretation

Table 1 presents the global performance of the visual-based and acoustic-based phone recognizers, with performance of the visual phone recognizer broken down into frontal and profile visual lip input features.

Results are significantly improved compared to [4], (54 % *vs* 60 %), but because the two databases have neither the same text material nor the same speaker, a strict comparison is not in order. The similar performances between [4] and this study may rather be interpreted as evidence of the method's robustness. In fact, as results using the lateral lip view are almost identical in the two studies (54 % [4] *vs* 56 %), much of the improvement (60 %) appears to be due to the use of the frontal lip view. This result agrees with the human lipreading experiments described in [15] as well as the digit recognition task detailed in [16].

	ASR	VSR = Tongue + Lips	
		Lateral	Frontal
P	83.9 %	56 %	**60 %**
Δ	0.7 %	0.9 %	0.9 %
D	1226	4776	4424
S	2389	7830	7036
I	1985	2695	2430
N	33637		

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

The performance of the visual phone recognizer can be analyzed using a confusion matrix as displayed at figure 5. As could be expected, it is phones with similar articulatory gestures (tongue and lips), {[p],[b],[m]}, {[k],[g],[ng]}, {[f],[v]}, {[s],[z]} and {[t],[d],[n]}, which are the most often confused by the system. Some of the vowel mismatches are quite "reasonable", such as [uh] (book) confused with [uw] (boot), and [iy] (beet) interpreted as [ih] (bit). The confusion of several phones with schwa [ah] can be explained by the well-known reduction phenomenon for vowels, or in other cases by the presence of a syllabic consonant, such as the [l] in "bottle". Diphthongs for which a tongue glide is involved are sometimes confused with one of their pure vowel components, for example [ey] (bait), [oy] (boy) and [ow] being matched with [ah], [iy] and [ao] (caught) respectively. The matrix also clearly shows an error occurring mainly on dental and alveolar sounds {[th],[dh]} (thin, then) and {[t],[d],[s],[sh]}. This is explained by the lack of information about the tongue tip (apex) in the ultrasound images, which is sometimes hidden by the acoustic shadow of the mandible. The relatively high number of insertions has a negative impact on the global performance, and the use of a statistical language model would certainly be helpful here. Finally, the predicted phonetic sequence is plagued by a large number of deletion errors. The phones which are most often deleted are very short ones such as the schwa [ah], as well phones corresponding to rapid articulatory gestures such as {[t]-[d]-[n]}. In fact, with a mean duration of 60 ms, the phone [t] is most often represented by fewer than two ultrasound frames with our current 30 Hz acquisition setup. A faster acquisition system is in the planning stages.

As the partitioning of phonetic space used is very fine (40 phonetic classes), our 60 % result is in fact pessimistic; it would no doubt be higher if some of the "reasonable" confusions mentioned, as well as mismatches due to incorrect phonetic labeling, were not considered "errors" in the performance computation. Too, such mismatches in the recognition stage need not necessarily lead to unintelligible synthesis. Some psychoacoustic effects and results provided by speech perception theory could potentially also be used to advantage. Thus, though as yet not perfect, our results are already promising enough to warrant investigating the feasibility of a phonetic vocoder driven by ultrasound and optical images of the tongue and lips.

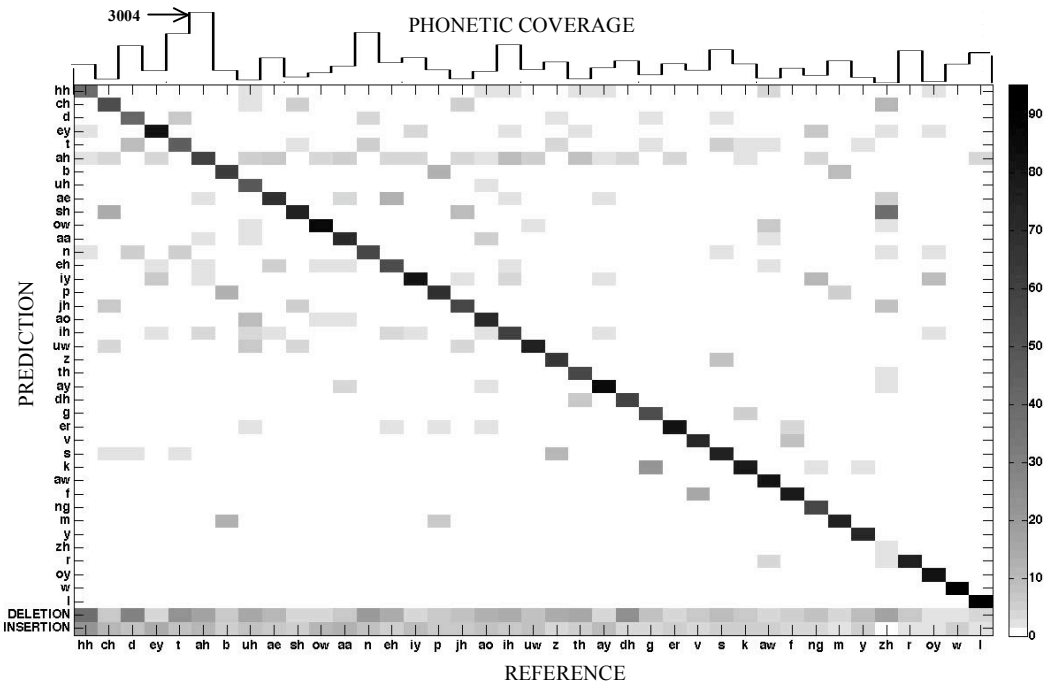

Figure 5: *Confusion matrix for phone recognition from ultrasound tongue sequences and frontal lip views. The color space map was chosen to emphasize the errors. Phone labels are in the TIMIT format. The histogram shows the number of occurrences of each phone in the database.*

4. Conclusions and perspectives

The visual phone recognizer is able to predict a 60 % correct phonetic target sequence from a continuous stream of video-only data. Applied to two different databases, with different textual materials and speakers (one male, one female), the proposed method appears robust, and could be a good starting point for phonetic vocoder driven only by visual observation of the voice organ. It is clear, however that the problem of phone insertion and deletion must be addressed more aggressively. The use of a language model and the acquisition of data at a higher rate are to be investigated in future work. We also intend to take into account possible asynchronies between articulators and compare the *feature fusion* strategy to a multistream HMM-based approach.

5. Acknowledgements

This work is supported by the French Department of Defense (DGA) and the French National Research Agency (ANR), 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] Nakajima, Y., Heracleous, P., Saruwatari, H., Shikano, K., "A Tissue-conductive Acoustic Sensor Applied in Speech Recognition for Privacy," Smart Objects & Ambient Intelligence Oc-EUSAI 2005, pp. 93-98, 2005.

[3] Denby, B., Oussar, Y., Dreyfus, G., Stone, M., "Prospects for a Silent Speech Interface Using Ultrasound Imaging," IEEE ICASSP, Toulouse, France, pp. I365- I368, 2006.

[4] 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.

[5] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Toward a Segmental Vocoder driven by Ultrasound and Optical Images of the Tongue and Lips," *submitted to these proceedings*.

[6] 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.

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

[8] Epstein, M., Stone, M., Pouplier, M., Parthasarathy, V., "Obtaining a palatal trace for ultrasound images," Proc. of Meeting of Acoustical Society of America, 2004.

[9] 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.

[10] 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/.

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

[12] 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," IEEE ICASSP, Honolulu, pp. I1245-I1248, 2007.

[13] Turk, M. A., Pentland, A. P., "Face Recognition Using Eigenfaces," IEEE Computer Soc. Conf. on Comp. Vision and Pat. Reco., Proc. CVPR, pp. 586-591, 1991.

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

[15] T. R. Jordan and S. M. Thomas, "Effects of horizontal viewing angle on visual and audiovisual speech recognition," in Journal of Experimental Psychology: Human Perception and Performance, vol. 27, no. 6, 2001, pp. 1386–1403.

[16] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," in Proceedings of the 8th IEEE Workshop on Multimedia Signal Processing (MMSP '06), pp. 24-28, Victoria, BC, Canada, October 2006.