Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface

Thomas Hueber [1,3], *Elie-Laurent Benaroya* [1], *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, laurent.benaroya@espci.fr, gerard.chollet@tsi.enst.fr, denby@ieee.org, gerard.dreyfus@espci.fr, mstone@umaryland.edu

Abstract

Recent improvements are presented for phonetic decoding of continuous-speech from ultrasound and optical observations of the tongue and lips in a silent speech interface application. In a new approach to this critical step, the visual streams are modeled by context-dependent multi-stream Hidden Markov Models (CD-MSHMM). Results are compared to a baseline system using context-independent modeling and a visual feature fusion strategy, with both systems evaluated on a one-hour, phonetically balanced English speech database. Tongue and lip images are coded using PCA-based feature extraction techniques. The uttered speech signal, also recorded, is used to initialize the training of the visual HMMs. Visual phonetic decoding performance is evaluated successively with and without the help of linguistic constraints introduced via a 2.5k-word decoding dictionary.

Index Terms: silent speech interface, visual speech recognition, multi-stream modeling

1. Introduction

Designing a device to allow speech communication without the necessity of vocalizing has become a challenge in its own right in the speech research community. This "Silent Speech Interface", or SSI, could be used to preserve the privacy of conversations, communicate in silence-restricted or high background noise environments, or for silent, hands-free data transmission during a security operation. Further applications are possible in the medical field, for example to assist laryngectomized patients, where the SSI would be used as an alternative to the electrolarynx; to oesophageal speech, which is difficult to master; or to tracheo-oesoephageal speech, which requires additional surgery. Different types of sensors can be envisaged in order to build an SSI. A speaker may for example produce airflow in his vocal tract and capture the resulting "murmur" with a stethoscopic microphone as in [1] and [2]. Other approaches, based on completely non-acoustic features have also been proposed, as for example in [3] where electromyographic electrodes placed on the speaker's face record muscular activity, or in [4] where magnets glued to the tongue and lips are tracked by sensors incorporated in a pair of eyeglasses. In our approach, articulator activity, mainly of the tongue and lips, is captured by a non-invasive multimodal imaging system composed of an ultrasound transducer placed beneath the chin and an optical camera in front of the lips [5].

In [6], we presented a framework for a phonetic vocoder driven exclusively by streams of visual observations, using an audio-visual unit dictionary that associates acoustic utterances with their visual phone equivalents. In the first stage of the system, the visuo-phonetic decoder finds the most likely phonetic targets for a given test sequence of visual data. These targets then constrain the selection in the dictionary of the sequence of units that best matches the input test data. In such a corpus-based approach, the quality of the synthesis depends strongly on the performance of the phonetic decoding stage, whose robustness must therefore be maximized. To that end, more sophisticated HMM-based modeling techniques have been recently tested. The two improvements presented and evaluated in the present paper are: the introduction of context-dependency in the modeling of the visual phones; and the use of a multi-stream approach to model jointly the ultrasound and the optical data streams. Systems derived from this approach will be compared to a baseline decoder, similar to that used in [6], which uses context-independent phonetic models and a feature fusion strategy. Because it is not *a priori* feasible to disambiguate all phonetic configurations only from tongue and lip observations, linguistic constraints can also be introduced to help the phonetic decoding, *via* for instance, a restriction on the allowed vocabulary. We therefore also evaluate our systems on both an unconstrained phonetic decoding task and on a more restricted one.

The development of the visuo-phonetic decoding baseline system is detailed in Section 2, where data acquisition and pre-processing, visual feature extraction techniques and evaluation protocols are described. Section 3 addresses the implementation and evaluation of the context-dependent visual phonetic decoder. In section 4, the multi-stream modeling approach is introduced and evaluated. Also in that section, the performance of the final system including both context dependent modeling and the multi-stream approach is discussed.

2. Baseline Visuo-Phonetic Decoder

2.1. Data acquisition and pre-processing

Ultrasound data is recorded using the Vocal Tract Visualization Lab HATS system [7]. In this setup, the transducer is locked in a fixed position beneath the chin, and the head immobilized. An acoustic standoff is used to allow mandible motion so that speech production is relatively undisturbed. Two standard video cameras record both profile and frontal views of the speaker's lips, and a microphone captures the uttered speech signal. The three video streams (two cameras plus ultrasound) and the audio signal are merged into the same video sequence using an analog video mixer, which limits the frame rate of the video data to 29.97 Hz (NTSC format). A typical image recorded by this acquisition system is shown in figure 1.

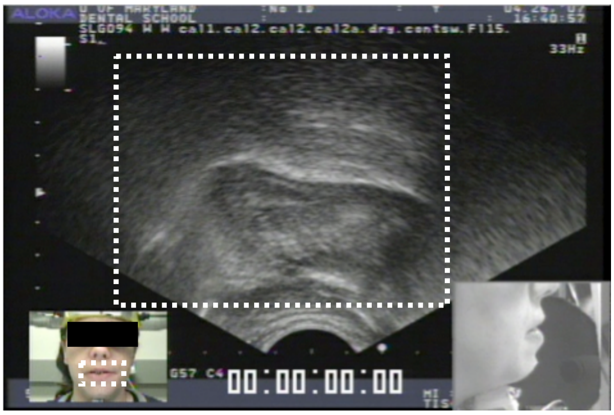

Figure 1: *An ultrasound vocal tract image in the mid-sagittal plan with embedded lip frontal and lateral view. Dashed white lines represent tongue and lip regions of interest.*

The text material, chosen for the purposes of diphone-based concatenative synthesis, is based on the first 1020 sentences of the CMU Arctic corpus [8], read by a native speaker of American English instructed to speak as neutrally as possible. After cleanup of the recordings, the database contains 61 minutes of speech contained in 109553 bitmap frames. Audio files are sampled at 16 kHz.

2.2. Visual feature extraction

Regions of interest (ROI) selected in ultrasound and optical images, as shown in figure 1, are first resized to 64x64 pixels. Speckle noise typical of ultrasound images is reduced using the anisotropic diffusion filter described in [9]. It is suggested in [10] that a frontal view of the lips provides more articulatory information than a profile; thus, although both are present in our database, we chose to use only the frontal view in this study. The "EigenTongues" [11] decomposition technique is subsequently used to encode each ultrasound frame. In this method, the vocal tract configuration is interpreted as a linear combination of standard configurations, the "EigenTongues", obtained by performing a Principal Component Analysis (PCA) on a phonetically balanced subset of frames. A similar technique is used to encode frontal images of the lips ("EigenLips"). The numbers 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 used on this database are 30 coefficients for each of the two streams. In order to be compatible with a more standard frame rate for speech analysis, the EigenTongues/EigenLips coefficient sequences are oversampled from 30 Hz to 100 Hz using linear interpolation. In this baseline system, EigenTongues/EigenLips coefficients, together with their first and second derivatives, are concatenated into a single "visual feature vector" in a feature fusion strategy. Dimensionality reduction techniques were not used in this analysis; their application is under study and will be addressed in a future work.

2.3. HMM-based modeling

The modeling of visual feature sequences by continuous HMMs requires their initial temporal decomposition at the phonetic level. As visual and audio modalities have been recorded synchronously, this initial segmentation can be derived from the labeling of the acoustic signal. This task is performed using a forced alignment procedure with an initial set of 40 acoustic HMMs trained on the acoustic component of the recorded database. The acoustic wave of each recorded sentence is parameterized by 12 Mel-frequency cepstral coefficients (MFCC) with their energies and first and second derivatives. In this study, all the procedures involving HMM manipulations are done using the HTK front-end [12]. After initialization, 40 left-to-right (monophones), 5-state (3 emitting states), continuous visual HMMs (with diagonal covariance matrices) are first trained separately using the standard Baum-Welch re-estimation algorithm. Then, embedded training, during which the number of Gaussians per state is incrementally increased, is used to refine the models and the temporal segmentation of the visual stream. In the testing stage, phonetic decoding is performed using the standard "Token Passing" algorithm, which finds the optimal path through an HMM network. Because some very important sources of information are missing in the visual data, such as nasality and the voiced/unvoiced flag, linguistic constraints can be introduced to help the phonetic decoding. With that in mind, we introduce two decoding scenarios. In the first, considered "unconstrained", the structure of the decoding network is a simple loop in which all phones loop back to each other. In the second, or "constrained" scenario, the phonetic decoder is forced to recognize words contained in the CMU Arctic sentences. In that case, the decoding network allows all possible word combinations which can be built from a 2.5k word dictionary. No statistical language model is used in the present study.

The 1020 sentences of the recorded database are divided into 34 lists of 30 sentences. In order to increase the statistical relevance of the speech recognizer performance, a jackknife (leave-one-out) technique [13] was employed, in which each list was used once as the test set while the other 33 lists composed the training set. Two test lists were however excluded from this jackknife procedure to be used as a validation set for the optimization of two "hyper" parameters: the model insertion penalty; and the number of Gaussians per state of the visual HMMs. For the baseline system, the optimal number of Gaussian per state was found to be 32.

For each phone class, a representative measure P of the recognizer performance is defined as:

$$P = \frac{N - D - S - I}{N} \qquad (1)$$

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. Section A of table 1 presents the performance of the baseline visuo-phonetic decoder in the two decoding scenarios.

3. Context-Dependent Modeling

Articulatory features such as those derived from the recorded images of the tongue and lips are naturally sensitive to context effects such as co-articulation and anticipation. Introducing context-dependency in the modeling of visual features sequences should therefore increase the robustness of the visuo-phonetic decoding. In this study, we propose to model visual triphones by adding information about left and right contexts to the phone models. Traditionally, triphone modeling presents several practical issues. Since many triphones have only a few occurrences in the training data, the accurate estimation of their corresponding HMM parameters is difficult. Also, many triphones may be missing in the training corpus, especially in a relatively small dataset such as the one used in this study. To overcome these issues and make visual triphone training viable, a tree-based state-tying strategy is adopted. Using the procedure in [14], a binary decision tree is constructed for each state of each phone, in order to cluster together all of the corresponding states of all of the associated triphones. The decision tree recursively partitions this pool of states by querying left/right contexts. States reaching the same leaf node are considered similar enough to be tied together.

The "yes - no" questions associated with the tree nodes are usually based on phonetic knowledge such as backness and height for vowels, place and manner of articulation for pulmonic consonants, etc. A typical question attached to a node of the decision tree might be, "Is the previous phone (left context) a bilabial consonant ?". However, as tongue and lip configurations are explicitly represented here, we propose to use a gesture-based approach to build the contextual questions. A feature set in which tongue body, tongue tip, and lip configurations are described explicitly, using the articulatory phonology theory introduced in [15], is used. With this description, the articulatory configuration corresponding to the phone [sh], for instance, would be characterized as a configuration where the lips are in a default "labial" position, the tongue tip in the "palato-alveolar" region, "tongue body" in the "palatal" region, etc. A typical contextual question on the decision tree built from this feature set would be, "Does the next phone (right context) require the tongue body moving to the palatal region?" However, although present in the feature set, no questions based on the glottal activity (which is meaningless in a silent speech context) or on the velum (which is not visible in an ultrasound image) was built.

To build the context-dependent visual phonetic decoder, a set of 40 visual HMMs (monophones) is first trained using the same procedure as for the baseline system. These monophone models are then cloned to initialize their corresponding untied triphones. As each training set of the jackknife procedure contains approximately 8500 distinct triphones with, on average, only 4 occurrences apiece, state tying appears to be essential. After the tree-based clustering procedure, the total of 25500 states (8500 x 3 emitting states) is reduced roughly to 1800 clusters (~7% of the original number of states). Tied-state models are then refined by incrementally increasing the number of Gaussian mixture components up to an optimal number which was found to be 4. Finally, models for unseen triphones are generated; decision trees are asked to find which combination of already trained state models is the most adapted to represent the context of a given unseen triphone. At the end of the training stage, a set of 67200 visual triphone models (all possible triphones and biphones built from a 40 element phone set) is available for decoding. As for the baseline system, decoder performance is evaluated on both unconstrained (free phonetic decoding) and constrained (using a 2.5k word dictionary) scenarios, as shown in sections A and B of Table 1. Compared to the baseline system, performance of the context-dependent system is significantly improved in both decoding scenarios, with improvements of 4.9% (unconstrained) and 3.2% (constrained) respectively. A more detailed analysis of the remaining decoding errors is given at the end of the next section.

4. Multi-Stream *vs.* Feature Fusion

As in audio-visual speech recognition, two approaches can be envisioned to integrate tongue and lip data streams in an HMM-based phonetic decoder: "feature fusion" which was adopted in the baseline system; and "(classifier) decision fusion". As described in [16], different strategies can be used to combine modalities at the classifier level. In this work, an "early integration" strategy based on state-synchronous Multi-Stream Hidden Markov Models (MSHMM, [17]) is used to model tongue and lip feature sequences. In a MSHMM, each stream has, for each state, its own Gaussian mixture and thus its own emission probability density function. Given a "tongue (T) and lips (L)" visual observation vector $o_t^{TL} = [o_t^T; o_t^L]$, the resulting emission likelihood b_j for state j is expressed as:

$$b_j(o_t^{TL}) = \prod_{S \in \{T,L\}} \left[\sum_{m_S=1}^{M_S} c_{jSm_S} N(o_t^S; \mu_{jSm_S}; \Sigma_{jSm_S}) \right]^{\lambda_S} \qquad (2)$$

where $N(o;\mu;\Sigma)$ is the value at o of a Gaussian mixture with mean μ and covariance Σ, M_S the number of mixture components and $\lambda_S = \{\lambda_T, \lambda_L\}$ are the weight parameters discussed below. In this equation, stream components are forced to be state-synchronous and thus asynchrony between tongue and lips movements, well described in [18], is not taken into account. However, since asynchrony is often correlated with phonetic context, the use of context-dependent models could potentially compensate this phenomenon. The combination of the stream likelihoods also requires the definition of the weight parameters λ_T and λ_L. Widely discussed in the context of audiovisual speech recognition (AVSR) [16], the estimation of stream exponents can be achieved either by measuring stream reliabilities using an SNR or a "degree of voicing" criterion for the audio modality, which is not possible here, or by maximizing system performance on a validation data set. In this initial test of multi-stream modeling of tongue and lip data, a very simple optimization procedure is adopted: only class-independent weights are used, and system performance is evaluated on a validation set for different pairs of weights, which we constrain to sum to one. As expected, the tongue carries the most important part of the accessible articulatory information, and the optimal values found for tongue and lip feature streams are $\lambda_T = 0.7$ and $\lambda_L = 0.3$.

In our procedure, a multi-stream phonetic decoder using context-independent models is first trained using the same procedure as for the baseline system. Its performance is shown in section C of Table 1. Compared to the baseline system, the multi-stream approach brings a 2% improvement in the unconstrained decoding scenario (with fewer substitution and insertion errors but more deletion errors), and a 3,7% improvement in the constrained one. Also, when the multi-stream approach is combined with context-dependent modeling, as in the "final" system whose performance is shown in section D of Table 1, the performance improvement is about 8% higher than that of the baseline system. While still not ideal, these results are nonetheless promising and demonstrate the relevance of the two new adopted strategies.

Quite naturally, most of the substitution errors are made on phones with similar tongue and lip gestures, such as

Table 1. *Performance of the different visuo-phonetic decoders. Δ is the 95% confidence interval, "CI", "CD", "Unconst.", and "Const." stand for context-independent and context dependent models, unconstrained and constrained decoding scenarios, respectively.*

	A		**B**		**C**		**D**	
	Baseline Decoder *CI – Feature Fusion*		**Context-dependent Decoder** *CD – Feature Fusion*		**Multi-Stream Decoder** *CI – MSHMM*		**Context-dependent & Multi-Stream Decoder** *CD - MSHMM*	
	Unconst.	**Const.**	**Unconst.**	**Const.**	**Unconst.**	**Const.**	**Unconst.**	**Const.**
P	57,7%	67,4%	62,6%	70,6%	59,5%	71,1%	65,6%	74,7%
Δ	1.0%	1.0%	1,0%	1,0%	1.0%	1,0%	1.0%	1.0%
D	6043	4666	3452	3196	7531	5174	4294	3964
S	7077	5398	7080	4799	5897	4157	6279	3613
I	1568	1270	2451	2210	658	696	1397	1232
N	34693	34693	34693	34693	34693	34693	34693	34693

{[p],[b],[m]}, {[t],[d],[n]}, {[f],[v]}, {[k],[g],[ng]}, {[ch],[jh]}, {[sh],[zh]} and {[th],[dh]}. In fact, if we consider these phone groups as equivalence classes, in which within-group confusions are not counted as errors, the performance of the final system in the unconstrained scenario can be further increased to 73,2% (78% for the constrained decoding scenario). Most of the remaining substitution errors are due to: vowels confused with the phone [ah], which is, in continuous speech, certainly a consequence of the vowel reduction effect; diphthongs matched sometimes with one of their vowel components; and dental and alveolar consonants, which are difficult to image with ultrasound because the apex (tongue tip) may be hidden by the acoustic shadow of mandible. Some of these mismatches in the phonetic decoding would not necessarily lead to unintelligible synthesis; some psychoacoustic effects could potentially also be used to advantage. The relatively high number of deletion and insertion errors, however, remains problematic, and will continue to be addressed in future work.

5. Conclusions

In order to improve the visuo-phonetic decoding stage of a planned ultrasound-based silent speech interface, the modeling of tongue and lips feature sequences using multi-stream and context-dependent HMMs has been proposed. On an open-vocabulary continuous speech decoding task, the system is able to correctly identify 65,6% of the phones from visual information only. When the vocabulary is limited to 2.5k words, the performance increases to 74,7%. Compared to a baseline system based on context-independent models and a feature fusion strategy, this new approach has led to a 8% absolute performance improvement. To reduce the remaining decoding errors (mainly deletions and insertions), the recording of visual data at a higher frame rate and the use of a statistical language model are currently under study.

6. 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.

7. References

[1] Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., "Non-audible murmur recognition", Proc. of Eurospeech, pp. 2601-2604, 2003.

[2] Tran V.-A., Bailly G., Lœvenbruck H., Jutten C., "Improvement to a NAM captured whisper-to-speech system", Interspeech, Brisbane, Australia, pp. 1465-1498, 2008.

[3] 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.

[4] Fagan, M.J., Ell, S.R, Gilbert, J.M., Sarrazin, E., Chapman, P.M., 2008. Development of a (silent) speech recognition system for patients following laryngectomy. Medical Engineering & Physics, vol. 30, issue 4, pp. 419-425.

[5] 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.

[6] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips", Interspeech, pp. 2028-2031, Brisbane, Australia, 2008.

[7] 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.

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

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

[10] Lucey, P., Potamianos, G., "Lipreading using profile versus frontal views", in Proc. IEEE MMSP'06, Canada, pp. 24-28, 2006.

[11] 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.

[12] 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/.

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

[14] Young, S. J., Odell, J. J., Woodland, P. C. "Tree-based state tying for high accuracy acoustic modeling", In Proc. of the Workshop on Human Language Technology, pp. 307-312, 1994.

[15] Browman, C. P., Goldstein, L., "Gestural specification using dynamically-defined articulatory structures", Journal of Phonetics, 18, pp. 299-320, 1990.

[16] Potamianos, G., Neti, C., Luettin, J., Matthews, I., "Audio-Visual Automatic Speech Recognition: An Overview". In: Issues in Visual and Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press, 2004.

[17] Bourlard, H., Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands", In Proc. ICSLP, pp. 426-429, 1996.

[18] Livescu, K., Glass, J., "Feature-based pronunciation modeling for speech recognition", in Proc. HLT/NAACL, 2004.