Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips

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

[1]Université Pierre et Marie Curie – Paris VI, 4 place Jussieu, 75252 Paris Cedex 05 France
[2]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
[3]Laboratoire Traitement et Communication de l'Information, Ecole Nationale Supérieure des Télécommunications (ENST-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

The article describes a video-only speech recognition system for a "silent speech interface" application, using ultrasound and optical images of the voice organ. A one-hour audio-visual speech corpus was phonetically labeled using an automatic speech alignment procedure and robust visual feature extraction techniques. HMM-based stochastic models were estimated separately on the visual and acoustic corpus. The performance of the visual speech recognition system is compared to a traditional acoustic-based recognizer.

Index Terms: speech recognition, audio-visual speech description, silent speech interface, machine learning

1. Introduction

In recent years, several systems using articulatory data to synthesize speech in real time have been described in the literature. These data may be derived from EMG/EPG measures [1], from a "non audible murmur microphone" signal (NAM [2]) or, in our case, from images of the voice organ [3]. Such a synthesizer, driven only by articulatory data, may be qualified as a "silent speech interface" (SSI), in that it could be used as an alternative to tracheo-oesophageal speech for laryngeal cancer patients, in situations where silence must be maintained, or for voice communication in noisy environments.

In [4], a static neural network was used to learn the "visuo-acoustic" mapping between the ultrasound tongue and optical lip images and a set of Line Spectrum Frequencies (LSF). That study demonstrated the relevance of visual features for describing the voice organ but permitted only LPC-based synthesis without an appropriate excitation signal.

Here, we propose visual speech recognition as a first step towards corpus-based silent speech synthesis, which furthermore allows the possibility of introducing linguistic constraints in our analysis. Our approach is based on building a corpus associating video-extracted visual feature sequences to phoneme labels. HMM-based stochastic models trained on this database are then used to predict target phonetic sequences. An overview of the system is given in Figure 1.

Section 2 of the article details data acquisition and ultrasound image preprocessing, while speech segmentation techniques appear in section 3. The visual feature extraction is discussed in section 4. Visual speech recognition procedures are presented in section 5 and our results detailed in section 6. A discussion of the results and some ideas for future improvements appear in section 7.

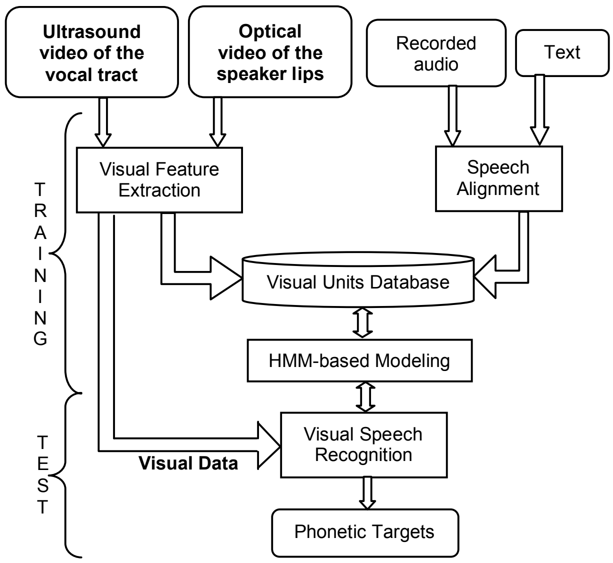

Figure 1: *An overview of the visual speech recognition system. Features derived from images, text and acoustic signals are combined together in an audio-visual speech corpus. HMM-based modeling method is used for speech recognition from video-only data.*

2. Data acquisition and preprocessing

An audio-visual database comprising video sequences of the voice organ together with the uttered speech signal was constructed using the Vocal Tract Visualization Lab HATS system [5]. This system is needed to fix the speaker's head and support the ultrasound transducer under the chin without disturbing speech. A lip profile image is embedded into the ultrasound image, as shown in figure 2.

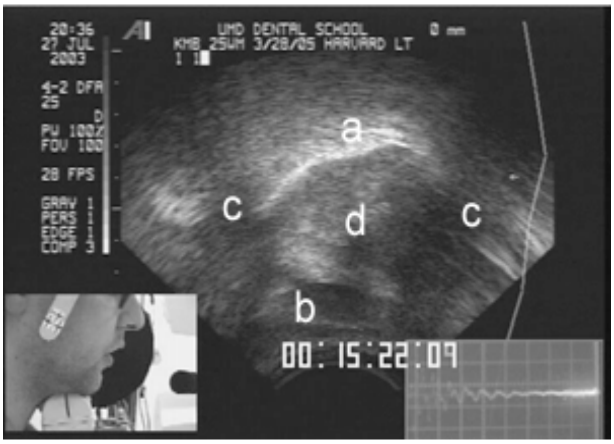

Figure 2: *Example of an ultrasound vocal tract image with embedded lip profile: (a) tongue surface; (b) hyoid bone; (c) hyoid & mandible acoustic shadows; (d) muscle, fat, connective tissue within the tongue.*

The recorded speech dataset consists of the 720 sentences (organized in 72 lists) of the IEEE/Harvard corpus [6] pronounced by a male native American English speaker. After cleaning the database, the resulting speech (43 minutes) was stored as 72473 JPEG frames and 720 WAV audio files sampled at 16000 Hz. The IEEE/Harvard base was initially chosen because all sentences have roughly equal intelligibility and approximately the same duration, grammatical structure and intonation across lists. They are furthermore constructed to preserve the mean frequencies of occurrence of segmental phonemes in the English language. The 1889 words of the IEEE/Harvard base were transcribed into phoneme sequences using the CMU[1] and British English[2] (BEEP) pronunciation dictionaries. The phonetic coverage of the sentences is shown in figure 3.

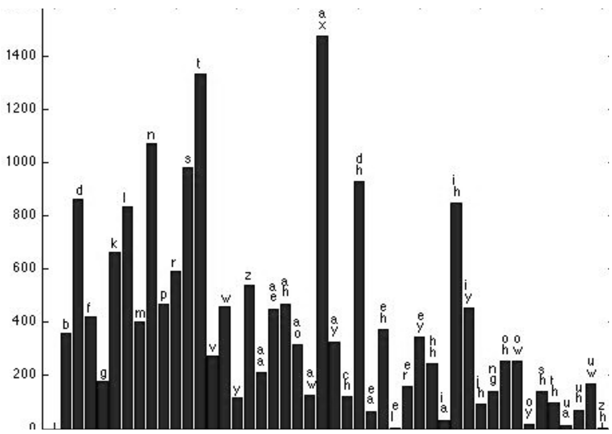

Figure 3: *Phonetic coverage of the IEEE/Harvard sentences (phone labels are written in the TIMIT format). The mean and standard deviation of the number of phone occurrences are 393.5 and 358.5, respectively.*

[1] www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

3. Segmental speech description

As the visual and audio streams are synchronized, the initial phonetic segmentation of the video sequences can be obtained from the temporal boundaries of the phonemes in the audio signal. The alignment procedure can thus be viewed as a simplified recognition task in which the phonetic sequence is already known. The HTK front-end [7] was used to accomplish this task. The speech acoustic signal is first parameterized using 12 Mel-frequency cepstral coefficients, with their normalized energies, deltas and accelerations. The transcribed multi-speaker DARPA TIMIT speech database [8] is then used to build initial HMM acoustic models. These 5-state, 16 mixture, left-to-right HMM models are finally applied to segment the audio stream of the corpus.

4. Visual feature extraction

The ultrasound images are first reduced to a polar region-of-interest grid delimited by the acoustic shadows (figure 2) of the hyoid bone and mandible. In [4], a PCA-based feature extraction approach, called "EigenTongues" in analogy to Turk and Pentland's "EigenFaces" for face recognition [9], interpreted the ultrasound image as a linear combination of standard vocal tract configurations, thus extracting more information from the images than a contour-based approach. In the present study, this method was improved by adapting the coding of the standard vocal tract configurations to a speech description context. Rather than using a random subset of frames to build the basis vectors of the "EigenTongue" decomposition, visual units from each phone class were picked to constitute the training database. This guarantees a better exploration of the possible vocal tract configurations, and tests showed that equivalent coding quality could be obtained with fewer input features than in the previous method. An analogous approach, called "EigenLips," was used to code the lip frames. The first three basis vectors of the "EigenTongue" and "EigenLip" decompositions are shown in figure 4. Finally, in order to improve the segmentation precision, visual feature sequences were oversampled from 30 Hz to 100 Hz, using linear interpolation.

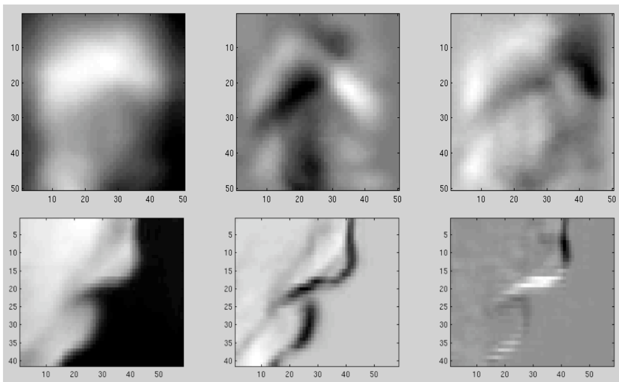

Figure 4: *The first three EigenTongues (top) / EigenLips (bottom), from left to right.*

5. Visual speech recognition

As our speech database is less than one hour long, and as shown in figure 3, actually has rather sparse phonetic coverage, the use of context-dependant models cannot be envisioned in this study. Rather, a set of 45 left-to-right, 5-state, continuous monophone HMM's is used to model the visual observation sequences of each phoneme class. Each visual observation is composed of 15 EigenTongues and 5 EigenLips with their delta and acceleration coefficients, centered and normalized. Once this initial set of models has been created and initialized, embedded training is performed, and the HMM models are incrementally refined by increasing the number of Gaussians per state to 32.

Visual speech recognition is performed using a Viterbi algorithm which finds the optimal path through the word model network, where word models are obtained by concatenating phone HMM models. As the experiment is intended to show the quality of the HMM-based modeling, no statistical language model is used in this study. Thus, speech recognition is constrained only by the use of a pronunciation dictionary built from the IEEE/Harvard sentences, containing in our case 2390 items (some words of the IEEE/Harvard corpus are transcribed with several pronunciations).

In order to increase the statistical relevance of the speech recognizer performance, a jackknife (leave-one-out) technique [10], in which each list of ten sentences was used once as the test set, was employed. For each phone class, a representative measure PA of the recognizer performance is defined as

$$PA = 100\frac{X}{N} = 100\frac{N-D-S-I}{N} \quad (1)$$

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. A 95% confidence interval Δ is computed from the Wilson formula [11]:

$$\Delta = 100\frac{\frac{t_\alpha}{N}\sqrt{1+4NX(1-X)}}{1+t_\alpha^2/N} \quad (2)$$

using $t_\alpha = 1.95$ and a normal approximation.

An identical procedure was used for traditional speech recognition based on acoustic features. As in visual speech recognition, a set of 45 context-independent, left-to-right, 5-state, 16-mixture, continuous monophone HMM's is estimated on each training pass. Because the goal of this study is not to achieve high accuracy on audio speech recognition, no more sophisticated modeling methods have been employed. In fact, adding or removing state transitions, tying parameters between models or forming some context-dependant models where possible could have refined these acoustic models. As such, the performance of our acoustic-based speech recognizer can be considered as a target for this database.

6. Results

Figure 5 illustrates qualitatively the performance of the visual speech recognizer on an example in which the predicted phone sequence is time aligned with the reference phonetic transcription via a dynamic programming based string alignment procedure. Correct predictions as well as errors are apparent.

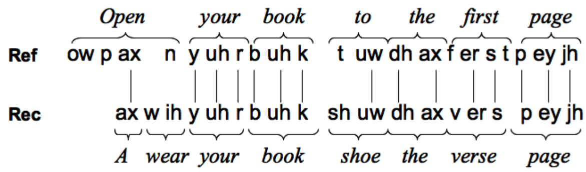

Figure 5: *Reference phonetic transcription (Ref) and predicted phonetic transcription (Rec) derived from visual features.*

The overall performance figures of video-only and audio-only speech recognition experiments are presented in Table 1. As the visual (VSR) and audio (ASR) speech recognizers share the same decoding dictionary and do not use language models, the accuracy of the visual HMM models as compared to ASR can be directly deduced from the table. Though as yet inadequate for synthesis purposes, the results are nonetheless quite promising.

Table 1. *Performance comparison of the visual (VSR) and acoustic-based (ASR) speech recognizers.*

Criterion	ASR	VSR
PA	71.0 %	54.5 %
Δ	1.3 %	1.4 %
D	874	2994
S	2485	4123
I	2101	1459
N	18874	

The high deletion error rate in visual speech recognition, defined as

$$d = 100\times\frac{D}{N} \quad (3)$$

may be due to the original video sampling rate of 30 Hz. Indeed, this rate makes the visualization of the vocal tract configuration difficult for very short phones, as illustrated in Table 2.

Table 2. *Relation between deletion error rate and mean phone duration. Illustration for phones having the first three highest/lowest deletion error rates.*

Phoneme	*d*	Mean Duration
Dh	37.3 %	0.05 s
T	19.3 %	0.09 s
ax	17.7 %	0.05 s
sh	4.0 %	0.17 s
uw	3.2 %	0.12 s
ey	1.6 %	0.16 s

A decomposition of our results into the different phoneme classes appears in Table 3. The recognition scores of plosives (*p, b*), fricatives (*f, v*) and nasal (*m,n*) phonemes show that labial movements are relatively well detected. Velar sounds (*ng, k, g*), formed by the tongue body and articulated near the soft palate, are also well recognized. However, vocal tract configurations corresponding to dental sounds (*th, dh*) and alveolar sounds (*s, sh, t, d*) are more difficult to detect. This can be explained by the lack of information about the relative position of the apex (tip of the tongue) and the teeth. Indeed, in the ultrasound images, the apex is hidden by the acoustic shadow of the mandible. Finally, the performance on vowel detection, which can theoretically be classified by how far forward and how high the tongue is in the mouth, is more difficult to interpret, and for some phonemes (*ah, uh*), the performance of our VSR system is quite low. It seems likely that context-independent HMM models used are not efficient enough to cope with the variability of these phones caused by the co-articulation phenomena.

Table 3. *Visual speech recognizer performance PA by phoneme, where Δ is the 95 % confidence interval and N the number of occurrences.*

Phone	Typical word	*PA* (in %)	*N*	Δ (in %)
zh	azure	0	1	NA
hh	**h**ay	9.7	256	7.2
ah	b**u**t	19.6	322	8.5
ch	**ch**oke	27.4	142	14.3
sh	**sh**e	32.9	149	14.7
uh	b**oo**k	34.2	114	16.9
jh	**j**oke	35.3	99	18.1
er	b**ir**d	41.9	203	13.3
ih	b**i**t	43.4	934	6.3
ae	b**a**t	47.4	449	9.1
z	**z**one	52	713	7.2
th	**th**in	52	98	19.1
dh	**th**en	53	915	6.4
y	**y**acht	53.5	114	17.7
d	**d**ay	54	995	6.1
eh	b**e**t	58	379	9.8
ax	**a**bout	58.5	1767	4.6
t	**t**ea	58.9	1733	4.6
b	**b**ee	59.3	440	9.1
uw	b**oo**t	59.8	249	11.9
n	**n**oon	60.6	1453	5
v	**v**an	60.7	349	10
ao	b**ou**ght	62	600	7.7
aa	b**o**tt	62	261	11.6
g	**g**ay	62.5	224	12.4
ey	b**ai**t	64	425	9
ow	b**oa**t	66.5	323	10.1
m	**m**om	68.5	524	7.9
ix	deb**i**t	68.7	32	29
f	**f**in	69.6	539	7.7
p	**p**ea	70.6	582	7.3
s	**s**ea	71.8	1131	5.2
ng	si**ng**	74.7	186	12.2
el	bott**le**	75	24	30.6
aw	b**ou**t	75.1	173	12.6
iy	b**ee**t	75.2	733	6.2
ay	b**i**te	76.2	425	8
r	**r**ay	82.5	1157	4.3
k	**k**ey	86.4	805	4.7
w	**w**ay	87.9	537	5.5
l	**l**ay	90	1121	3.5
oy	b**oy**	91.8	49	14.7

7. Conclusions and perspectives

The ability to extract discrete phones from continuous physiological data of the voice organ will be an important step in the design of a silent speech interface. In this article, promising, relevant performance measures have demonstrated the feasibility of phone recognition from ultrasound images of the tongue and optical images of the lips.

At present, the single target phonetic sequence derived from the visual features cannot directly be used to drive the research of acoustic segments in the corpus. The single target will have to be enlarged to a lattice of phonetic targets through which a data-driven unit search of the corpus can correct the stochastic model prediction errors. It would also be desirable to provide improved visual features. The use of optical flow based techniques [12], for example, is currently under study in order to model the movement of the visible articulators. The visual speech recognition will furthermore have to be validated on a larger dictionary with a robust language model, or as a limiting case, without any dictionary. Finally, the construction of a larger database, with a higher video sample rate and an additional front view of the speaker's face, is foreseen.

8. Acknowledgements

The authors would like to acknowledge useful discussions with Leila Zouari, Elsa Angelini, Yacine Oussar, and Pierre Roussel. This work is supported by the French Department of Defense (DGA) and the French National Research Agency (ANR).

9. 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, 2006, pp. I365- I368.

[4] 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, 2007, to appear.

[5] 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.

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

[7] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., The HTK Book, Revised for HTK Version 3.3, September 2005, *http://htk.eng.cam.ac.uk/.*

[8] 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.

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

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

[11] Hogg, R.V., Tanis, E.A., Probability and statistical inference (6th ed.), Upper Saddle River, NJ: Prentice Hall.

[12] Angelini, E., Gerard, O., "Review of myocardial motion estimation methods from optical flow tracking on ultrasound data," IEEE EMBS Annual International Conference, New York, NY, USA, pp.1537-1540, 2006.