Recognition and Real Time Performances of a Lightweight Ultrasound Based Silent Speech Interface Employing a Language Model

Jun Cai [1, 2], *Bruce Denby* [1, 2], *Pierre Roussel* [2], *Gérard Dreyfus* [2], *Lise Crevier-Buchman* [3]

[1] Université Pierre et Marie Curie, Paris, France
[2] SIGMA Laboratory, ESPCI ParisTech, CNRS-UMR 7084, Paris, France
[3] Laboratoire de Phonétique et Phonologie, CNRS-UMR 7018, Paris, France

{Jun.Cai, Denby}@ieee.org, {Pierre.Roussel, Gerard.Dreyfus}@espci.fr,
lise.buchman@numericable.fr

Abstract

The work presents advances in the implementation of an ultrasound based silent speech interface system. Use of a portable acquisition device, a visual speech recognizer system with a language model, and real time tests with the Julius system are described. Experiments with two types of visual feature extraction are also presented. Results show that good recognition and real time performance can be obtained with a portable silent speech interface employing a language model.

Index Terms: silent speech interface, visual speech recognition, vocal tract imaging, ultrasound imaging

1. Introduction

A silent speech interface (SSI) is intended to enable speech communication in the absence of an intelligible acoustic signal [1]. Several experimental SSI systems have been developed using a variety of different sensors [1]. The REVOIX project at the Sigma Laboratory in Paris is building an SSI meant to restore the voices of speech-impaired individuals in real-time. The technique chosen for REVOIX is to drive a recognizer-synthesizer system using ultrasound and video images of the tongue and lips. The REVOIX SSI thus consists of three modules operating sequentially: (1) an acquisition module to record simultaneous ultrasound and visual images of the vocal tract; (2) a word-level visual speech recognizer that uses Hidden Markov Models trained on features extracted from these images (HTK toolkit [7]), rather than from acoustic features; and (3) a speech synthesizer. To be genuinely useful, such a device will ultimately have to be lightweight, have good recognition and synthesis performance, and operate in real time.

In this report, we build upon the groundwork laid in earlier research [2-6] by:

- Introducing a new, portable acquisition system;
- Comparing different types of visual feature extraction;
- Introducing the use of a language model to improve the recognition accuracy;
- Experimenting with a real time implementation of the recognition using the Julius system.

Our results show that it is possible to obtain good recognition and real time performance using a portable SSI system employing a language model.

The visual speech acquisition system and the acquired corpora are described in Section 2 and 3. In Section 4, two visual speech feature extraction techniques, namely the EigenTongues/EigenLips and the Discrete Cosine Transform (DCT), are presented. The experimental results are given in Section 5. Conclusions are drawn in Section 6.

2. Visual Speech Data Acquisition

The multimodal speech data acquisition system is shown in Figure 1. It is comprised of a lightweight, adjustable helmet (Figure 1(a)) housing an 8MC4 microconvex ultrasound probe (opening angle: 140°, frequency range: 4-8 MHz) for tongue imaging; a CMOS industrial camera for imaging the lips; and a lapel microphone for audio recording if desired (Figure 1(b)) (the audio signals are not used in the REVOIX SSI). An infrared illumination and an infrared filter are affixed to the camera to make video acquisition independent of ambient lighting conditions.

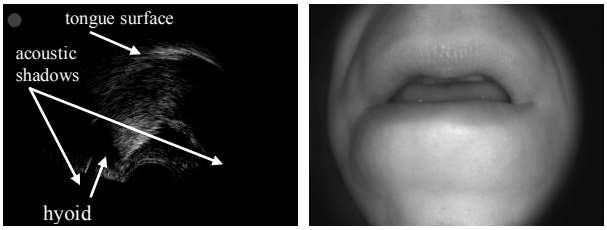
(a) (b)

Figure 1: *Lightweight helmet, with ultrasound probe, infrared camera, and lapel microphone.*

The ultrasound system used is the lightweight t3000™ from Terason. The ultrasound and video imaging devices are controlled by a stand-alone, easy to operate, dedicated graphical software interface called Ultraspeech [9]. Ultraspeech uses a multithread programming technique to allow synchronous acquisition of the two image streams at their respective maximum frame rates, along with an audio signal. At an ultrasound focal distance of 7 cm, appropriate for tongue visualization, the system records, simultaneously and synchronously, the ultrasound stream at 60 fps (image resolution of 320×240 pixels), the video stream at 60 fps (image resolution of 640×480 pixels), and the audio signal (16 KHz, 16 bits). A typical pair of synchronous ultrasound and video images of the tongue and lips is shown in Figures 2(a-b). Because a large amount of memory is need for buffering the acquired image streams, the maximal duration of a recorded utterance is limited to 8s by Ultraspeech. In REVOIX, an ordinary laptop PC is used to run the software, and the entire SSI system can be fit into a small carrying case.

An acquisition protocol was developed for recording the visual speech data. The training corpus was organized into lists of 50 sentences. To avoid speaker fatigue, the acquisition was split into several sessions separated by intervals of at least 24 hours. Within each session, several lists were recorded.

After each list, the subject takes a short break, without removing the helmet, and drink water with a straw to hydrate the mouth. After every two lists, the calibration of tongue and lips images is checked, and if necessary, the sensors be adjusted. Additional ultrasound gel is also placed on the surface of the ultrasound probe after every two lists to maintain good ultrasound imaging quality.



(a) (b)

Figure 2: *Ultrasound tongue, and infrared lip images*

Since the visual speech features are extracted from the ultrasound and video images, maintaining the positioning consistency of the ultrasound probe and video camera across all training sessions is critical to the performance of the SSI. Recording a large amount of multimodal speech data for training the recognition models requires multiple sessions spaced in time, thus necessitating an inter-session re-calibration mechanism to maintain fixed positions of the sensors used, and possible readjustment between sessions. The Ultraspeech graphical user interface includes a module which allows the user to interactively re-calibrate ultrasound and video images before data acquisition to maintain positioning consistency, as described in [5,6]. During re-calibration, the subject can thus adjust the positions of the ultrasound probe and video camera to match the current images to pre-recorded target reference images. In this way, the positioning consistency of the sensors can be maintained.

3. Visual Speech Corpora

Data acquisition for the visual speech recognizer in the REVOIX project remains a time and labor-intensive task. The results presented here are for a single speaker. The TIMIT [10] text was used for constructing the training. All 2342 sentences of the TIMIT text were uttered once by a male native English speaker (Speaker BD), silently and in the non-verbalized punctuation (NVP) manner. During the utterances, visual speech data were recorded using the system described in Section 2. This recorded visual speech corpus, which we shall call the Visual TIMIT training corpus, is suitable for training the English phone models since the TIMIT text set is phonetically balanced.

Two additional corpora were constructed for our performance evaluation. The first was built by using a subset of WSJ0 5,000-word test set [11]. One hundred short sentences were selected from the 330 WSJ0 5,000-word test sentences, and read by the Speaker BD in the same way as that for recording the Visual TIMIT corpus. Hereafter we refer to this test set as the Visual WSJ0 5k test corpus. The second one, which we shall call the Visual Gigaword 20k test corpus, consists of 200 sentences extracted in late February 2011 from the four English newswire sources used in the English Gigaword corpus [12]. The sentences were selected such that the words contained in them are included in the English Gigaword 20k vocabulary. The texts of these two test corpora can be consulted at [13].

4. Visual Speech Feature Extraction

Two different visual feature extraction techniques were implemented and compared, as described below.

4.1. EigenTongues/EigenLips Approaches

The "EigenTongues" and "EigenLips"approaches [14] were first used to extract visual speech features from both the ultrasound and video images. In the "EigenTongues" technique, each ultrasound image is projected onto the feature space of "EigenTongues", which can be seen as the space of standard vocal tract configurations obtained after a Principal Components Analysis (PCA) using a set of 500 randomly selected frames. The "EigenLips" decomposition was used to encode video images of the lips. Before performing these decompositions, ultrasound and video regions of interest were first resized to 64×64 pixels. The numbers of projections onto the sets of EigenTongues and EigenLips used for coding were determined by keeping the eigenvectors carrying at least 80% of the variance of the training set. In this work, 30 coefficients for each of the two streams were extracted for representing each combined video frame. Specifically, denote the EigenTongues as a $p_T \times k_T$ matrix $\boldsymbol{V}_T$, where p_T is the number of dimensions of each eigenvector, k_T (=30) is the number of eigenvectors for projection. A resized tongue image $\boldsymbol{T}_n$ can be decomposed as:

$$\boldsymbol{T}_n = \boldsymbol{V}_T \times \boldsymbol{F}_{Tn} + \boldsymbol{\mu}_T + \boldsymbol{\varepsilon}_{Tn} \quad (1)$$

where $\boldsymbol{F}_{Tn}$ is the k_T -dimensional score vector of this image frame determined by PCA; $\boldsymbol{\mu}_T$ is the p_T -dimensional mean vector; and $\boldsymbol{\varepsilon}_{Tn}$ is the PCA decomposition deviation vector of $\boldsymbol{T}_n$. Similarly, for a resized video image $\boldsymbol{L}_n$ of lips, the PCA decomposition can be expressed as:

$$\boldsymbol{L}_n = \boldsymbol{V}_L \times \boldsymbol{F}_{Ln} + \boldsymbol{\mu}_L + \boldsymbol{\varepsilon}_{Ln} \quad (2)$$

Using a "feature fusion strategy", the PCA scores of tongue and lips images were concatenated into a single vector, along with their first and second derivatives, resulting in visual speech feature vectors with 180 components.

4.2. DCT-based Approach

The discrete cosine transform (DCT) is a technique widely used for lossy image compression. DCT provides a representation of the frequency content of the transformed image. It has a strong "energy compaction" property: most of the signal information tends to be concentrated in a few low-frequency components of the DCT [15]. Based on the assumption that the most relevant information in the tongue and lips images is carried by components with low spatial frequencies, the DCT technique can be adopted to extract visual speech features. This was the second technique we used.

As an example, for a resized tongue image represented by a matrix $\boldsymbol{A}$ of size $M \times M$, the two-dimensional DCT is computed as:

$$\boldsymbol{D}_{ij} = a_i a_j \sum_{m=0}^{M-1} \sum_{n=0}^{M-1} A_{mn} \boldsymbol{cos} \frac{\pi(2m+1)i}{2M} \boldsymbol{cos} \frac{\pi(2n+1)j}{2M}, \quad 0 \le i,j \le M-1 \quad (3)$$

where

$$a_i, a_j = \begin{cases} \frac{1}{\sqrt{M}}, & i,j = 0 \\ \sqrt{\frac{2}{M}}, & 1 \le i,j \le M-1 \end{cases}$$

To be consistent to the EigenTongues/EigenLips feature extraction in terms of the feature vector dimension, the 30 lowest frequency components were extracted from the DCT of a tongue image to form the tongue DCT vector. Similarly,

another 30-dimensional vector was extracted from the DCT of the corresponding lip image. These two vectors were then concatenated, along with their first and second derivatives, resulting in DCT feature vectors again of 180 components.

5. Experimental Results and Analyses

5.1. Continuous Visual Speech Recognition on PCA Features

The HTK 3.4 toolkit was used to train the visual speech HMM models on the PCA features of the Visual TIMIT training corpus. These HMMs were built in the form of cross-word triphones in order to capture the coarticulatory effects both within words and across words in the continuous visual speech. An empirical study was conducted to vary the number of Gaussians in each GMM from 2 to 16. An 8-Gaussian GMM for each HMM state was found accurate enough to model cross-word triphones in this research.

Continuous visual speech recognition was carried out on the two test sets by using HTK 3.4. For recognizing the Visual WSJ0 5k test set, since no suitable trigram model was available, only the WSJ0 5k NVP bigram [11] language model was used. For the recognition on the Visual Gigaword 20k test set, however, a Gigaword 20k trigram model was also used. The recognition accuracies were evaluated on both the word-level and phone-level. The results are shown in Table 1. As a simple word-loop bigram model was employed in earlier SSI work [6], it was again included here in order to extend the comparison of different language models, and to illustrate the impact of an appropriate language model on recognition performance for our application.

Table 1. *Recognition accuracy on the two test sets, using PCA features and different bigram LMs.*

Test Set + Language Model	Recognition Accuracy (%)	
	Word Level	Phone Level
Visual WSJ0 5k + WSJ0 5k NVP bigram	79.08	90.21
Visual WSJ0 5k + Word-loop bigram	43.50	77.87
Visual Gigaword 20k + Gigaword 20k bigram	69.05	86.54
Visual Gigaword 20k + Gigaword 20k trigram	77.53	89.82
Visual Gigaword 20k + Word-loop bigram	12.06	70.72

It is observed that, with word-loop bigram models, the accuracy was much lower, on both test sets, than that obtained using task specific bigrams or trigrams, which impose a strong domain-specific constraint on the search space. The explanation for this is that the probability distributions of words and word strings in the word-loop bigram model are quite different from those in the test set. As an example, two test sentences from our two test sets are shown in Tables 2 and 3, where the word-level outputs relevant to different bigram models are shown. One may remark that some non-grammatical word strings, such as "two be peers" and "wee half tear", have occurred in the recognition results from the word-loop bigrams. The word-loop bigram performance on Visual Gigaword 20k is even further reduced compared to WSJ0 5k because the 4-fold increase in vocabulary greatly expands the number of possible incorrect word choices.

Though the word level recognition accuracy on both test sets was less than 80%, on the phone level, the accuracy was higher. This demonstrates that visual speech data can be well represented by PCA features, and that using tied-state cross-word triphone HMMs and a bigram model does allow visual speech to be well decoded at the phone-level.

Table 2. *Word recognition output and original transcript of a visual speech utterance in Visual WSJ0 5k test corpus.*

Original Text		we're going to be bidders said a top official of a major oil company
Recognized Text	WSJ0 5k NVP bigrams	we're going to be peers said top official of a much oil company
	Word-loop bigrams	we'll going two be peers san top official of up h. up oil company

Table 3. *Word recognition output and original transcript of a visual speech utterance in Visual Gigaword 20k test corpus.*

Original Text		now we have the earthquake which is going to knock the numbers around a bit
Recognized Text	Gigaword 20k bigrams	now we have the earthquake which is going to knock numbers around a pin
	Gigaword 20k trigram	now we have the earthquake which is going to knock numbers around a bit
	Word-loop bigrams	now wee half tear earthquake witch is going tune knocked numbers around up inn

From Table 2, it can also be seen that even when the domain-specific WSJ0 5k NVP bigram model was used, "bidders" was misrecognized as "peers", and "major" as "much". A reasonable explanation for this is that since no information about the larynx height and the tongue tip position can be acquired in the ultrasound and video images, the vocal tract configuration of phones such as "d" and "jh", whose pronunciations rely heavily on the movement of larynx and/or tongue tip, is not well represented by the visual speech data.

5.2. Continuous Visual Speech Recognition on DCT Features

DCT-based features were also used to perform the visual speech recognition on the two test corpora. The HMM models had the same structure and same complexity as those in Section 5.1, but were trained by using the DCT features of the Visual TIMIT training corpus. This recognition was also performed by using HTK 3.4. Results are presented in Table 4.

Table 4. *Recognition accuracy by using DCT features on the Visual WSJ0 5k test set.*

Test Set + Language Model	Recognition Accuracy (%)	
	Word Level	Phone Level
Visual WSJ0 5k + WSJ0 5k NVP bigram	84.16	93.44
Visual Gigaword 20k + Gigaword 20k bigram	76.91	90.22
Visual Gigaword 20k + Gigaword 20k trigram	86.07	93.81

It can be seen that, for each test set, using the same language model, a significant increase (about 5%) of the recognition accuracy was introduced by using DCT features.

This indicates that for the visual speech recognition in the current portable REVOIX SSI setup, the DCT-based feature extraction technique performs better than the PCA-based technique in terms of recognition accuracy.

5.3. Real Time Factor of the Recognition using HTK

It was found in our experiments that by using DCT features, the recognition system also ran faster than the system using PCA features. In Table 5, the real-time factors of the recognition using PCA-based features are compared with those of the recognition using DCT-based features. As the HMM model complexity was the same for both types of features, it is hypothesized that the DCT features, being somewhat "sharper", led to a more compact set of paths in the Viterbi search procedure. The faster execution of the trigram compared to bigram is believed to be due to the trigram constraint being applied over a longer window, again limiting the search space.

Table 5. *Real-time factor for recognition on the test sets, using HTK 3.4. vs. using Julius 4.1.5*

Test Set + Language Model	Real-time Factor (CPU Time/Audio Time)	
	PCA-based	DCT-based
HTK 3.4		
Visual WSJ0 5k + WSJ0 5k NVP bigram	3.61	0.56
Visual Gigaword 20k + Gigaword 20k bigram	26.22	3.56
Visual Gigaword 20k + Gigaword 20k trigram	7.96	1.93
Julius 4.1.5		
Visual WSJ0 5k + WSJ0 5k NVP bigram	0.46	0.42
Visual Gigaword 20k + Gigaword 20k bigram	0.66	0.59
Visual Gigaword 20k + Gigaword 20k trigram	0.65	0.58

5.4. Continuous Visual Speech Recognition Using Julius

To improve the real-time performance of our recognizer, the Julius 4.1.5 system was also used to perform the recognition on the test sets. The triphone HMM models based on PCA features and DCT features were employed directly in the recognition experiments using Julius. The recognition accuracy was similar to that derived by using HTK 3.4, however, with a real-time factor less than one. The real-time factors for recognition on both test sets are also listed in Table 5. The difference between bigram and trigram speeds observed with HTK is not present here due to the 2 pass search method used in Julius.

6. Conclusions and Discussions

Our results show that, for the speaker tested here, ultrasound and video streams of the tongue and lips recorded during speech production can effectively be used to drive a continuous visual speech recognizer. The EigenTongues/ EigenLips and DCT-based approaches appear to be appropriate for constructing visual speech features with high precision. With the current REVOIX setup, the DCT-based approach performed better than EigenTongues/ EigenLips in both recognition accuracy and real-time performance. A set of tied-state cross-word triphone HMMs can be trained on the visual speech corpus, and by using the HMMs and a well-defined domain-specific bigram or trigram model, high recognition accuracy can be achieved, both at phone-level and word-level.

These results imply that the recognized text could be used as input to a subsequent speech synthesizer in an SSI to generate intelligible speech. By implementing the visual speech recognizer in the Julius system, word-level recognition can be performed in nearly real-time, with only a small loss in recognition accuracy.

7. Acknowledgements

This work was supported by the French National Research Agency (ANR) under contract numbers ANR-09-ETEC-005-01 and ANR-09-ETEC-005-02 REVOIX. The authors wish to acknowledge the contribution of Thomas Hueber GIPSA-Lab, Grenoble, France, who designed the data acquisition system, and laid the groundwork for a number of the experimental and analytical techniques used in this work.

8. References

[1] Denby, B., Schultz, T., Honda, K., et al., "Silent Speech Interfaces", Speech Communication, 52(4): 270-287, Apr. 2010.

[2] Young, S., Evermann, G., Gales, M., et al., "The HTK Book", Online: http://htk.eng.cam.ac.uk/docs/docs.shtml, accessed on 15 Apr. 2010.

[3] Hueber, T., Benaroya, E. L., Chollet, G., et al., "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips", Speech Communication, 52(4): 288-300, Apr. 2010.

[4] Hueber, T., Chollet, G., Denby, B., et al., "Visuo-Phonetic Decoding Using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface", Proc. INTERSPEECH 2009: 640-643, UK, Sept. 2009.

[5] Hueber, T., Chollet, G., Denby, B., et al., "Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips", Proc. INTERSPEECH 2008: 2028-2031, Australia, Sept. 2008.

[6] Hueber, T., Chollet, G., Denby, B., et al., "Phone Recognition from Ultrasound and Optical Video Sequences for a Silent Speech Interface", Proc. INTERSPEECH 2008: 2032-2035, Australia, Sept. 2008.

[7] Hueber, T., "Reconstitution de la Parole par Imagerie Ultrasonore et Vidéo de l'Appareil Vocal: vers Une Communication Parlée Silencieuse", Doctorate Thesis, Université Pierre et Marie Curie, Dec. 2009.

[8] Lee, A., Kawahara, T. and Shikano, K., "Julius – An Open Source Real-time Large Vocabulary Recognition Engine", Proc. EUROSPEECH 2001: 1691-1694, Denmark, Sept. 2001.

[9] Hueber, T., Chollet, G., Denby, B., et al., "Acquisition of Ultrasound, Video and Acoustic Speech Data for a Silent-speech Interface Application", Proc. International Seminar on Speech Production: 365-369, Strasbourg, France, Dec. 2008.

[10] Garofolo, J. S., Lamel, L. F., Fisher, W. M., et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Online: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1, accessed on 22 Mar. 2011.

[11] Garofalo, J., Graff, D., Paul, D., et al., "CSR-I (WSJ0) Complete", Online: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S6A, accessed on 22 Mar. 2011.

[12] Graff, D., Cieri, C., "English Gigaword", Online: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05, accessed on 22 Mar. 2011.

[13] Cai, J., Denby, B., "Transcripts of English Test Corpora", Online: http://ftp.espci.fr/shadow/SSI_Test/, accessed on 22 Mar. 2011.

[14] Hueber, T., Aversano, G., Chollet, G., et al., "Eigentongue Feature Extraction for an Ultrasound-based Silent Speech Interface", Proc. of ICASSP 2007: 1245-1248, Honolulu, USA, Apr. 2007.

[15] Rao, K. R. and Yip, P., "Discrete Cosine Transform: Algorithms, Advantages, Applications", Academic Press, Boston, 1990.