A VISUAL SPEECH RECOGNITION SYSTEM FOR AN ULTRASOUND-BASED SILENT SPEECH INTERFACE

Jun Cai[a,b], *Thomas Hueber*[c], *Bruce Denby*[a,b], *Elie-Laurent Benaroya*[d], *Gérard Chollet*[e], *Pierre Roussel*[b], *Gérard Dreyfus*[b] *& Lise Crevier-Buchman*[f]

[a]Université Pierre et Marie Curie, Paris, France;
[b]SIGMA Laboratory, ESPCI ParisTech, CNRS-UMR 7084, Paris, France;
[c]Département Parole & Cognition, GIPSA-Lab, CNRS-UMR 5216, Grenoble, France;
[d]Agro ParisTech & INRA, CNRS-UMR 518, Paris, France;
[e]Laboratoire Traitement et Communication de l'Information, Telecom ParisTech, Paris, France;
[f]Laboratoire de Phonétique et Phonologie, CNRS-UMR 7018, Paris, France
denby@ieee.org; Jun.Cai@ieee.org; thomas.hueber@gmail.com;
laurent.benaroya@gmail.com; chollet@telecom-paristech.fr; Pierre.Roussel@espci.fr;
Gerard.Dreyfus@espci.fr; lise.buchman@numericable.fr

ABSTRACT

The development of a continuous visual speech recognizer for a silent speech interface has been investigated using a visual speech corpus of ultrasound and video images of the tongue and lips. By using high-speed visual data and tied-state cross-word triphone HMMs, and including syntactic information via domain-specific language models, word-level recognition accuracy as high as 72% was achieved on visual speech. Using the Julius system, it was also found that the recognition should be possible in nearly real-time.

Keywords: silent speech interface, visual speech recognition, vocal tract ultrasound imaging

1. INTRODUCTION

The silent speech interface (SSI) is an emerging technology intended to enable speech communication in the absence of an intelligible acoustic signal. A number of experimental SSI systems have been developed, using different approaches to acquire sensor data from the elements of the human speech production process [1]. The REVOIX project at the Sigma Laboratory aims to build an SSI to restore the original voices of speech-impaired individuals, ultimately in real-time. Based on previous research and development of an SSI prototype [2, 4, 6-8], the fundamental mechanism chosen in REVOIX is to restore the speech using a recognizer-synthesizer system driven by ultrasound and video images of the tongue and lips. The image sequence is acquired by the image acquisition module during speech production; and is then transcribed into word-level text by the visual speech recognizer; which in turn, is passed to the speech synthesizer to generate a speech signal.

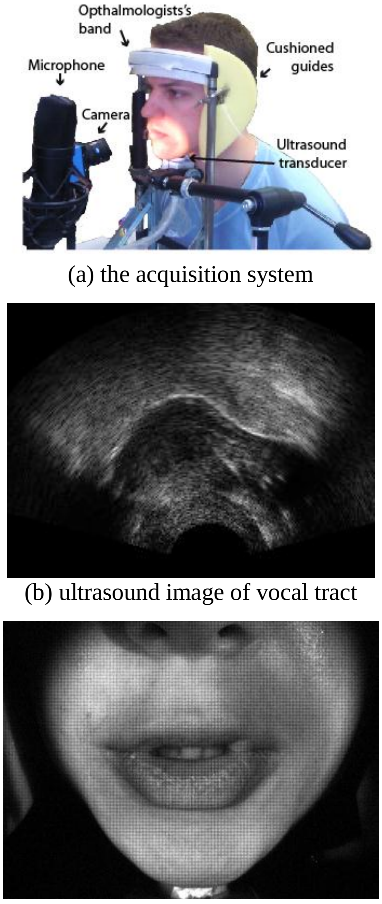
This research was focused on building a high-performance continuous visual speech recognition system within the framework of the REVOIX SSI. The HTK toolkit [12] was used to develop our HMM-based speech recognizer. Word-level recognition was performed with a view to driving a text-to-speech system for synthesizing continuous speech. To obtain a high accuracy, language models have been introduced to investigate how a well defined language model can contribute to the accuracy. To achieve good real-time performance, the two-pass large vocabulary continuous speech decoder Julius [11] was tested to implement the visual speech recognizer.

The visual speech acquisition system and the acquired corpus are described in Section 2. In Section 3 and 4, the methods for building the HMMs and language models are presented, respectively. The experimental results are given in Section 5. Conclusions are drawn in Section 6 with some discussions about visual speech recognition.

2. SPEECH DATA ACQUISITION AND CORPUS

The acquisition system for recording multimodal speech data is shown in Figure 1(a). The ultrasound transducer is placed beneath the chin via an articulated arm, while a video camera is placed ahead of the lips. The two imaging devices are controlled by a stand-alone and simple-to-operate graphical software interface called Ultraspeech [5]. At an ultrasound focal distance of 7cm, which is appropriate for tongue visualization, the system is used to record, 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 1(b) and 1(c).

Figure 1: Illustration of a dual video frame.



(a) the acquisition system

(b) ultrasound image of vocal tract

(c) video image of frontal view of lips

The first 1110 of the 1132 sentences contained in the CMU ARCTIC corpus [9], were each uttered once by a female native English speaker in the non-verbalized punctuation (NVP) manner. The acquisition was split into 10 sessions. An interactive inter-session re-calibration mechanism [5] was employed to maintain the positioning accuracy of the video sensors across all sessions.

The visual speech corpus is quite small. Testing the recognition system only once on a small part of the corpus would thus not be appropriate to evaluate the recognition performance in a statistical sense. To deal with this, a jackknife resampling [10] was performed by dividing the visual speech corpus into 37 subsets of 30 sentences. Each subset was used once for test while the others formed the corresponding training set, resulting in 37 jackknife tests.

3. VISUAL SPEECH FEATURES

The “EigenTongues” approach [3] was used to extract visual speech features from the ultrasound images. 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) of a subset of typical frames. In order to guarantee a good exploration of the possible vocal tract configurations, this subset is constructed so as to be phonetically balanced. A similar “EigenLips” decomposition was used to encode video images of the lips. Before performing these decompositions, ultrasound and video regions of interest were resized to 64×64 pixel size. The numbers of projections onto the set of EigenTongues/EigenLips used for coding were determined by keeping the eigenvectors to carry at least 80% the variance of the training data; typical values used on this corpus are 30 coefficients for each visual modality. The tongue and lip features were concatenated into a single vector, along with their first and second derivatives, resulting in visual feature vectors of 180 components.

4. VISUAL SPEECH RECOGNIZER

4.1. HMM modeling

The HTK 3.4 toolkit [12] was used to train the visual speech HMM models. In each jackknife test, the visual speech features and the phoneme transcripts of the training set were first used, via the HTK tools, to train 1-Gaussian monophone HMM models. Context-dependent phoneme transcripts were then created for the training sentences to train the triphone models. In this work, the 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.

4.2. Language modeling

Because it is not a priori feasible to disambiguate all phonetic configurations from tongue and lip observations alone (with no information on larynx activity or velum position), linguistic constraints must be introduced to facilitate the visual speech decoding. In our previous work, these constraints were introduced via an allowed vocabulary. Here, we add more linguistic information via a statistical language model, built up at word-level.

For each jackknife test, a dedicated stochastic bigram model, hereafter called the “ARCTIC” bigrams, was built in the NVP manner using the original source texts used to create the CMU ARCTIC database. These texts consist of 37 documents, most of which are stories of the early 20th century writer Jack London. The 37 texts [10] were preprocessed to segment them into sentences by treating periods, semicolons, and exclamation and question marks as separators between sentences. The lexicon of the CMU ARCTIC corpus contains 2,271 words. Using the 37 original texts, all sentences which contain only words found within the CMU ARCTIC vocabulary were extracted, excluding the sentences for test. This produced, for each jackknife test, a corpus of 29,827 sentences which was used to build the ARCTIC bigrams. These bigram models are closed-vocabulary, domain-specific language models, and are suitable for jackknife tests on the recorded visual corpus, although the vocabulary remains quite small.

To enlarge the scope of the vocabulary, a second NVP bigram model was also built for each jackknife test. The vocabulary of this bigram model consists of the union of the 2,271 words in the CMU ARCTIC lexicon and the 5,000 most-frequent words in the CMU ARCTIC source texts. All 76,501 sentences composed of only the words within this vocabulary were then extracted from the 37 source texts. By excluding the test sentences, these were then used to train our “ARCTIC-5k” bigram models. The ARCTIC-5k bigrams thus contain many words and word sequences not found in the CMU ARCTIC lexicon. Compared to the ARCTIC bigrams, the ARCTIC-5k bigrams impose a less restricted word-level constraint on the Viterbi search.

Since a simple word-loop bigram model was adopted in our previous SSI work [2], it was again included in this research to make a comparison of different bigram models, as well as to see the impact of the use of a well-defined LM on our recognition performance. In this word-loop model, any word pair in the CMU ARCTIC vocabulary is allowed with equal likelihood. This bigram model we call hereafter “ARCTIC word-loop”.

4.3. Using Julius for real-time performance

The jackknife tests have been carried out to evaluate the recognition accuracy of the visual speech recognizer. For the *i*th (1 ⩽ *i* ⩽ 37) jackknife test, the *i*th subset of the corpus was used as the test set, while the other 36 subsets forming the training set for building the triphone models. 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 our triphones.

Recognition was performed using the three bigram models described in Section 4.2 in each of the jackknife tests. The Viterbi word recognizer HVite of HTK was used to perform the word-level recognition. Both word-level and phone-level recognition accuracy were evaluated for each jackknife test. The overall results are shown in Tables 1 and 2.

Table 1: Word recognition accuracy of the 37 Jackknife tests.

Bigram	Recognition Accuracy (%)	
	Mean	Std.
ARCTIC bigrams	72.93	5.99
ARCTIC-5k bigrams	72.20	5.63
ARCTIC word-loop	56.90	7.18

Table 2: Phone recognition accuracy of the 37 Jackknife tests.

Bigram	Recognition Accuracy (%)	
	Mean	Std.
ARCTIC bigrams	83.39	3.68
ARCTIC-5k bigrams	84.09	3.17
ARCTIC word-loop	81.67	3.54

Table 3: Word recognition output of a visual speech utterance.

Original Text		our mr howison will call upon you at your hotel
Recognized Text	ARCTIC bigrams	i our mr howison will call upon you in your hotel
	ARCTIC-5k bigrams	i our mr howison will call upon you an’ you’re hotel
	ARCTIC word-loop	i our mr allow is in when call upon you in you’re owe tell

It is observed that by using the ARCTIC bigrams and ARCTIC-5k bigrams, each of which imposes a strong domain-specific constraint on the search space, the average word-level recognition accuracies were above 72.00%. With the ARCTIC word-loop bigram model, however, the accuracy was significantly lower. The explanation for this is that the probability distributions of words and word strings in the ARCTIC word-loop bigram model are quite different from those of the CMU ARCTIC text. The ARCTIC word-loop contains many word strings which do not occur either in the CMU ARCTIC text or in normal everyday speech.

As an example, the 19th sentence in subset 31 is shown in Table 3, where the word-level transcription outputs relevant to different bigram models are listed. Some non-grammatical word strings such as “allow is in” and “you’re owe tell” have occurred in the recognition results from the ARCTIC word-loop.

At the phone-level, the accuracy derived using even the ARCTIC word-loop, however, is above 80%, which is consistent with what was obtained in [2], even though in [2], by using the ARCTIC word-loop and 4-Gaussian word-internal triphones, a word recognition accuracy of 56.9% and a phone recognition accuracy of 83.3% were obtained, higher than the counterpart results in Table 1 and 2. The recognition outputs are quite similar to the original text at the phone-level; this demonstrates that using tied-state cross-word triphone HMMs and a bigram language model does allow visual speech to be decoded well at the phone-level.

4.4. Visual speech recognition using Julius

During the jackknife tests, the HVite recognizer required more than 10 times real-time to decode visual speech. To evaluate the “real-time” performance of our recognizer, the Julius system was also tested to perform the recognition in the jackknife tests. The triphone HMM models and the ARCTIC bigrams were employed directly in the recognition experiments using Julius. An average word recognition accuracy of 83% was obtained, and a visual speech utterance of *t* seconds required only about 0.90*t* seconds on average to complete the word recognition, on a 2.00 GHz Intel Core2 Duo Processor E4400 PC with 2GB of RAM.

5. CONCLUSIONS AND PERSPECTIVES

Our results show that, at least for the speaker tested here, ultrasound and video streams of the tongue and lips recorded during speech production can be used to drive a continuous visual speech recognizer effectively. The “EigenTongues” and “EigenLips” approaches appear to be appropriate for construct-ing visual speech features with high precision. 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 model, good 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 with the Julius system, word-level recognition can be performed in nearly real-time, with only a small loss in recognition accuracy. Since the real-time performance of the Julius system would not be significantly deteriorated by using a trigram model, it may be possible to use a domain-specific trigram LM in the Julius to further improve the recognition accuracy.

6. 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.

7. REFERENCES

[1] Denby, B., Schultz, T., Honda, K., et al. 2010. Silent speech interfaces. *Speech Communication* 52(4), 270-287.

[2] Hueber, T. 2009. *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.

[3] Hueber, T., Aversano, G., Chollet, G. et al. 2007. Eigentongue feature extraction for an ultrasound-based silent speech interface. *Proc. ICASSP* Honolulu, USA, 1245-1248.

[4] Hueber, T., Benaroya, E.L., Chollet, G. et al. 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication* 52(4), 288-300.

[5] Hueber, T., Chollet, G., Denby, B. et al. 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. International Seminar on Speech Production* Strasbourg, France, 365-369.

[6] Hueber, T., Chollet, G., Denby, B. et al. 2008. Phone recognition from ultrasound and optical video sequences for a silent speech interface. *Proc. Interspeech* Australia, 2032-2035.

[7] Hueber, T., Chollet, G., Denby, B. et al. 2008. Towards a segmental vocoder driven by ultrasound and optical images of the tongue and lips. *Proc. Interspeech* Australia, 2028-2031.

[8] Hueber, T., Chollet, G., Denby, B. et al. 2009. Visuo-phonetic decoding using multi-stream and context-dependent models for an ultrasound-based silent speech interface. *Proc. Interspeech* UK, 640-643.

[9] Kominek, J., Black, A. 2004. The CMU arctic speech databases. *Proc. 5th ISCA Speech Synthesis Workshop* Pittsburgh, 223-224.

[10] Kominek, J., Black, A. 2010. CMU ARCTIC Databases for Speech Synthesis. Online: *http://festvox.org/cmu_arctic/cmu_arctic_report.pdf*, accessed on 5 Oct. 2010.

[11] Lee, A., Kawahara, T., Shikano, K. 2001. Julius – An open source real-time large vocabulary recognition engine. *Proc. Eurospeech* Denmark, 1691-1694.

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