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

Jun Cai [1], *Thomas Hueber* [2], *Bruce Denby* [3, 1], *Elie-Laurent Benaroya* [4], *Gérard Chollet* [5], *Victoria-M. Florescu* [1], *Pierre Roussel* [1], *Gérard Dreyfus* [1], *Lise Crevier-Buchman* [6]

[1] SIGMA Laboratory, ESPCI ParisTech, CNRS-UMR 7084, Paris, France
[2] Département Parole & Cognition, GIPSA-Lab, CNRS-UMR 5216, Grenoble, France
[3] Université Pierre et Marie Curie, Paris, France
[4] Agro ParisTech & INRA, CNRS-UMR 518, Paris, France
[5] Laboratoire Traitement et Communication de l'Information, Telecom ParisTech, Paris, France
[6] Laboratoire de Phonétique et Phonologie, CNRS-UMR 7018, Paris, France

Michael Cai 1/3/11 18:15
Mis en forme: Français

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 linguistic information via a domain-specific language model, word-level recognition accuracy as high as 85% was achieved on visual speech. Using the Julius system, it was also found that the recognition should be possible in nearly real-time.

Index Terms— Silent speech interface, visual speech recognition, vocal tract imaging, 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 [2-6], 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 REVOIX SSI consists of three functional modules operating sequentially: the image sequence is acquired by the (1) image acquisition module during speech production; and is then transcribed into word-level text by the (2) visual speech recognizer; which in turn, is passed to the (3) speech synthesizer to generate a speech signal. The usability of such an SSI system depends on many factors, the most important of which, for the speech recognizer, are the recognition accuracy and execution speed.

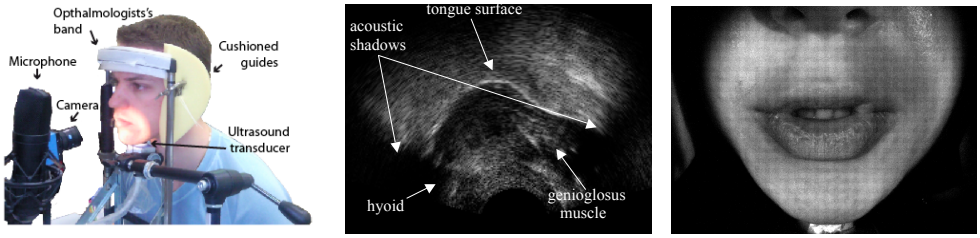
This research was focused on building a high-performance continuous visual speech recognition system within the framework of the REVOIX SSI. The HTK toolkit [7] was used to develop our HMM-based speech recognizer. Word-level visual speech recognition was performed with a view to driving a text-to-speech system for the generation of synthesized continuous speech.

To obtain a high accuracy, the HMMs were built in the form of cross-word tied-state triphone models. Language models have furthermore been introduced to investigate how a well defined language model can contribute to the recognition accuracy. To achieve good real-time performance, the two-pass large vocabulary continuous speech decoder Julius [8] was also 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 forehead of the subject rests upon an opthalmological stand, while the ultrasound transducer is placed beneath the chin via an articulated arm (a video camera is also placed ahead of the lips). The ultrasound system is the lightweight t3000™ developed by Terason, using an 8MC4 microconvex transducer (opening angle: 140°, frequency range: 4-8 MHz). Lip video is obtained with a 60 fps CCD industrial camera from The Imaging Source. The two imaging devices are controlled by a stand-alone, simple to operate, and dedicated graphical software interface called Ultraspeech [9] . Ultraspeech uses



(a) the acquisition system (b) ultrasound image of vocal tract (c) video image of frontal view of lips

Figure 1 Illustration of a dual video frame

a multithread programming technique to allow synchronous acquisition of the two video streams at their respective maximum frame rates, along with the audio signal. More specifically, at an ultrasound focal distance of 7 cm, 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).

The first 1110 of the 1132 sentences contained in the CMU ARCTIC corpus [10], were each uttered once by a female native English speaker in the non-verbalized punctuation (NVP) manner. To prevent speaker fatigue, the acquisition was split into 10 sessions spaced at an interval of at least 24 hours. An interactive inter-session re-calibration mechanism [9] was employed to maintain the positioning accuracy of the video sensors across all sessions.

Although the CMU ARCTIC text is in some sense phonetically balanced, the visual speech corpus itself is quite small. Testing the recognition system only once on a small part of the corpus would thus not be appropriate to evaluate the visual speech recognition performance in a statistical sense. To do so, a jackknife resampling [11] 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 FEATURE REPRESENTATION

The "EigenTongues" approach [12] was used to extract visual speech features from the ultrasound images. In this 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) 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. For both visual modalities, the number of projections onto the set of EigenTongues/EigenLips used for coding was set to 30, by empirically evaluating the quality of the image reconstructed from its first few components. Using a "feature fusion strategy", tongue and lip features were concatenated into a single visual feature vector, along with their first and second derivatives, resulting in vectors of 180 components.

4. DESIGN OF THE VISUAL SPEECH RECOGNIZER

4.1. HMM Modeling

The HTK 3.4 toolkit [7] was used to train the visual speech HMM models. Based on the CMU Pronouncing Dictionary, in which 39 phonemes are used, all sentences in the CMU ARCTIC corpus were encoded into phoneme sequences. In each jackknife test (see Section 5.1), the visual speech features and the phoneme transcripts of the training set were first used, via the HTK tools, to train 1-Gaussian HMM models of 40 monophones (including "silence"). Context-dependent phoneme transcripts were then created for the training sentences in order to train the triphone models.

In this work, the context-dependent HMMs had 3-state, left-to-right topology. 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. Since the number of model parameters increases dramatically using cross-word triphones, phonetic trees were used to perform parameter-sharing between the triphones, in order to handle the trainability issue. Thus, for each jackknife test, a set of tied-state cross-word triphone HMMs could be built. For our visual speech HMMs, the typical number of tied-states was 760, while the number of physical triphones was 3,771.

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

Thomas Hueber 2/3/11 17:40
Commentaire: This was the case before we checked that 30 components carry 80% of the variance (p 74 of my dissertation). So, we can replace this sentence by something like "The numbers of projections onto the set of EigenTongues/EigenLips used for coding are determined by keeping the eigenvectors carrying at least 80% of the variance of the training set; typical values used on this database are 30 coefficients for each of the two streams"

Thomas Hueber 2/3/11 17:37
Commentaire: I think that this paragraph has nothing to do in this section … it should be placed in the methodology or in the experimental results are presented.

Michael Cai 1/3/11 18:27
Supprimé: 11

Gérard Dreyfus 1/3/11 16:50
Supprimé: representative

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.

As the CMU ARCTIC sentences were extracted from out-of-copyright texts which are over 70 years old, contemporary English language models such as the CSR LM-1 and Gigaword LM [13] are inappropriate for recognizing the CMU ARCTIC utterances. Therefore, a domain-specific language model must be constructed.

For each jackknife test, a 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 [14] 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 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 model. 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 [6], it was again included in this research in order 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 to Improve Real-time Performance

In order to assess the feasibility of real-time recognition performance, the two-pass large vocabulary continuous speech decoder Julius [8] was tested to implement the visual speech recognizer. Sophisticated search techniques are incorporated in the Julius system, with the result that it can perform almost real-time *acoustic* speech decoding on a contemporary PC on a 60k-word vocabulary dictation. In our task, the CMU ARCTIC vocabulary contains only 2,771 words; therefore we might expect that a Julius-based *visual* speech recognizer could work in nearly real-time for our SSI tests.

5. EXPERIMENTAL RESULTS

5.1. Assessment of Recognition Accuracy by Jackknife Tests with Different Bigrams

The jackknife tests have been carried out to evaluate the recognition accuracy of the visual speech recognizer. For the ith ($1 \le i \le 37$) jackknife test, the ith 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. After recognition, the word transcription output was labeled into phoneme sequences using the HTK tool HLEd, based on the pronunciation dictionary. Both word-level and phone-level recognition accuracy were evaluated for each jackknife. 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

It is observed that by using the ARCTIC bigrams, which imposes a strong domain-specific constraint on the search space, the average word-level recognition accuracy was 72.93%. With the ARCTIC word-loop bigram model, the accuracy was 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,

Thomas Hueber 2/3/11 17:43
Commentaire: What do you mean by "real-time" … do we have to wait at least the end of the word ?

Michael Cai 1/3/11 18:28
Supprimé: 12

Gérard Dreyfus 1/3/11 17:42
Commentaire: There is a presentation problem here. This is a first time jacknife tests are mentioned in the paper, and they are actually defined in section 5.1. A part of the first paragraph of section 5.1 should be transferred here, I think. And the title of section 5.1 should be something like "assessement of recognition accuracy by jacknife tests for different bigrams"

Michael Cai 23/2/11 17:54
Supprimé: A…13…857 …This …is . … [1]

Michael Cai 1/3/11 18:16
Supprimé: by Using

Michael Cai 1/3/11 18:24
Supprimé: Although the CMU ARCTIC text is in some sense phonetically balanced, the visual speech corpus itself is quite small. Testing the recognition system only once on a small part of the corpus would thus not … [2]

Thomas Hueber 2/3/11 17:45
Commentaire: I think you can remove … [3]

Michael Cai 23/2/11 18:01
Supprimé: 85… … [4]

Michael Cai 23/2/11 18:00
Supprimé: . …These … [5]

Michael Cai 23/2/11 20:34
Supprimé: 79… … [6]

Thomas Hueber 2/3/11 17:47
Commentaire: We had 61.6% with La … [7]

Michael Cai 23/2/11 18:01
Supprimé: 57

Michael Cai 23/2/11 18:01
Supprimé: 7

Michael Cai 23/2/11 18:02
Supprimé: 91… … [8]

Michael Cai 23/2/11 20:34
Supprimé: 87… … [9]

Thomas Hueber 2/3/11 17:48
Commentaire: We had 83.3% with I … [10]

Michael Cai 23/2/11 18:02
Supprimé: 82

Michael Cai 23/2/11 18:02
Supprimé: 4

Michael Cai 23/2/11 18:02
Supprimé: 85…other two …s…these … [11]

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.

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

At the phone-level, the accuracy derived using even the ARCTIC word-loop, however, is above 80%, which is also consistent with what was obtained in [6]. It is clear that 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 model does allow visual speech to be decoded well at the phone-level.

5.2. 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 recognition accuracy of 83% was obtained, and a visual speech utterance of t seconds required only about $0.90t$ seconds on average to complete the word-level recognition, on a 2.00 GHz Intel Core2 Duo Processor E4400 PC with 2GB of RAM.

6. 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. 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 in 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 Julius to further improve the recognition accuracy.

7. ACKNOWLEDGEMENT

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.

8. REFERENCES

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

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

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

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

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

[6] T. Hueber, "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.

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

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

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

[10] J. Kominek, and A. Black, "The CMU Arctic Speech Databases," *in Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, pp. 223-224, June 2004.

[11] B. Efron, "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," *Biometrika*, 68(3), pp. 589-599, 1981.

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

[13] K. Vertanen, "Speech Recognition Training Recipes," *Online: http://www.keithv.com/software/*, accessed on 5 Oct. 2010.

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

Gérard Dreyfus 1/3/11 17:30
Supprimé: the

Thomas Hueber 2/3/11 17:51
Commentaire: I think that this table is useless. Adding a confusion matrix will be much much better, especially for ICPhS which is a phonetic conference ...

Thomas Hueber 2/3/11 17:53
Supprimé: The "EigenTongues" and "EigenLips" approaches appear to be appropriate for constructing visual speech features with high precision.

Unknown
Supprimé:

Michael Cai 23/2/11 20:38
Supprimé: high

Michael Cai 1/3/11 18:27
Supprimé: 11

Michael Cai 1/3/11 18:27
Supprimé: 12

Michael Cai 1/3/11 18:27
Supprimé: 13

Michael Cai 1/3/11 18:27
Supprimé: [14] B. Efron, "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," *Biometrika*, 68(3), pp. 589-599, 1981.