Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2009) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips

Thomas Hueber [a,b,*], Elie-Laurent Benaroya [a], Gérard Chollet [b], Bruce Denby [c,a], Gérard Dreyfus [a], Maureen Stone [d]

[a] 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
[b] CNRS-LTCI, Telecom ParisTech, 46 rue Barrault, 75634 Paris, Cedex 13, France
[c] Université Pierre et Marie Curie, 4 place Jussieu, 75252 Paris, Cedex 05, France
[d] Vocal Tract Visualization Lab, University of Maryland Dental School, 650 W. Baltimore Street, Baltimore, MD 21201, USA

Received 5 December 2008; received in revised form 15 October 2009; accepted 19 November 2009

Abstract

This article presents a segmental vocoder driven by ultrasound and optical images (standard CCD camera) of the tongue and lips for a "silent speech interface" application, usable either by a laryngectomized patient or for silent communication. The system is built around an audio–visual dictionary which associates visual to acoustic observations for each phonetic class. Visual features are extracted from ultrasound images of the tongue and from video images of the lips using a PCA-based image coding technique. Visual observations of each phonetic class are modeled by continuous HMMs. The system then combines a phone recognition stage with corpus-based synthesis. In the recognition stage, the visual HMMs are used to identify phonetic targets in a sequence of visual features. In the synthesis stage, these phonetic targets constrain the dictionary search for the sequence of diphones that maximizes similarity to the input test data in the visual space, subject to a concatenation cost in the acoustic domain. A prosody-template is extracted from the training corpus, and the final speech waveform is generated using "Harmonic plus Noise Model" concatenative synthesis techniques. Experimental results are based on an audiovisual database containing 1 h of continuous speech from each of two speakers.

Keywords: Silent speech; Ultrasound; Corpus-based speech synthesis; Visual phone recognition

1. Introduction

In recent years, the design of a device allowing speech communication without vocalization of the sound has emerged as a new research domain. This "Silent Speech Interface", or SSI, targets several applications. The device could for example be used for voice communication in a silence-restricted environment, either to preserve privacy in a public place, or for hands-free data transmission during a military or security operation. Based only on non-acoustic information, the SSI also enables voice communication even when speech is masked by background noise. The second main targeted application is medical: for assistance to a laryngectomized patient, using the SSI as an alternative to the electrolarynx; to oesophageal speech, which is difficult to master; or to tracheo-oesoephageal speech, which requires an additional surgery.

Different approaches and solutions have been adopted and described in the literature to make silent communication possible. For some applications, such as private communication by non-laryngectomized people, "absolute silence" may not always be necessary. A small airflow in the vocal tract can produce a "murmur" which is captured

* Corresponding author. Address: ESPCI ParisTech, Laboratoire d'Electronique, 10 rue Vauquelin, 75005 Paris, France. Tel.: +33 1 40 79 44 66; fax: +33 1 47 07 13 93.
E-mail address: hueber@ieee.org (T. Hueber).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.11.004

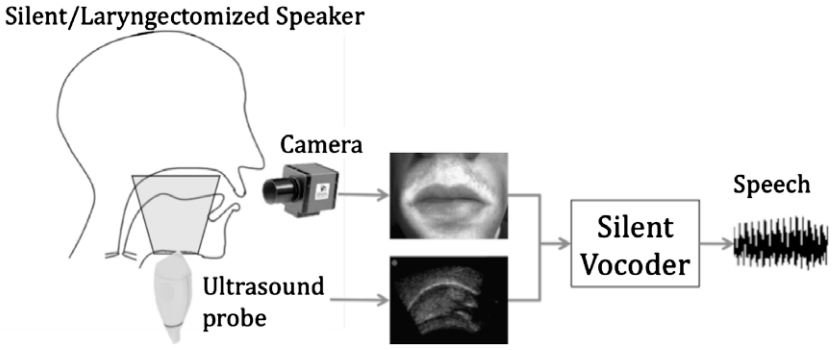

Fig. 1. Ultrasound-based silent speech interface (schematic).

by a stethoscopic microphone, as in (Heracleous et al., 2005; Tran et al., 2008). In other applications, the SSI must be able to recover an intelligible speech signal exclusively from a stream of non-acoustic features. These features measure the activity of the voice organ during the production of silent speech and can be derived from different measuring techniques. In (Maier-Hein et al., 2005), for example, several surface EMG electrodes placed on the speaker's face record myoelectric signals. In (Jorgensen et al., 2003), an electro-palatogram device (EPG) is combined with the same type of EMG surface electrodes, placed below the jaw in order to record the activity of the larynx and sublingual areas together with tongue-palate contacts. In (Fagan et al., 2008), magnets are glued to the tongue and lips and tracked by magnetic sensors incorporated in a pair of eyeglasses.

In our approach, a portion of the vocal tract is imaged using an ultrasound device coupled with a standard optical camera as shown in Fig. 1. This multimodal imaging system focuses mainly on tongue and lip visualization, even if some other vocal tract structures can also be observed. Because of its non-invasive property, clinical safety and good time resolution, ultrasound is well adapted to vocal tract imaging and analysis. Furthermore, since laptop-based high performance ultrasound imaging systems are available today,[1] a wearable real-time SSI system with an embedded ultrasound transducer and camera, can reasonably be envisioned.

The problem of speech synthesis from the analysis of articulatory motion is usually addressed by the use of a two- or three-dimensional articulatory model of the vocal tract (as in Maeda (1990), Sinder et al. (1997), and Birkholz and Jackèl (2003)). However, the state of the art in high-quality speech synthesis uses a segmental approach (or HMM-based methods, as discussed later), in which the speech waveform is obtained by concatenating acoustic speech segments. In our work, we propose to drive a segmental vocoder by visual observations of the articulators. In what follows, the terms "visual information" and "visual observations" are taken to refer both to the ultrasound and optical images.

Our approach integrates a phone recognition stage with a corpus-based synthesis system. As the first (off-line) step, an "audio–visual" dictionary is built from a training set containing units which associate the acoustic and visual realizations of each phone; this dictionary will later be used in the synthesis step. Given a test sequence of visual information only, the vocoder generates the speech waveform in three stages:

(a) An HMM-based decoder predicts a sequence of phonetic targets from the given set of visual features.
(b) A unit selection algorithm, driven by this prediction, searches in the dictionary the optimal sequence of audio–visual units that best matches the input test data.
(c) A speech waveform is generated by concatenating the acoustic segments for all selected units. Since there is no glottal activity, recovering an "acceptable" prosody from "silent data" is an issue, and prosodic transformations of the synthesized speech waveform are needed. These transformations are achieved using "Harmonic plus Noise Model" (HNM) (Stylianou et al., 1997) coding and synthesis techniques.

An overview of the recognition/synthesis system is given in Fig. 2.

Our study is based on a two-speaker (one male, one female, native speakers of American English), audiovisual database containing 1 h (per speaker) of ultrasound and optical images of the tongue and lips, recorded along with the acoustic signal. We point out that all experiments are performed here using uttered speech. The aim of the paper is to study the performance of our SSI on these databases; it will of course ultimately need to be evaluated on truly "silent" speech, which may be a non-trivial issue.

The second section of the article describes the data acquisition protocol and database content. The third section details visual feature extraction techniques, while the fourth describes the implementation of the visual phone recognizer and evaluates its performance for different levels of linguistic constraints. Finally, the unit selection algorithm, waveform

[1] For example, the Terason T3000 portable ultrasound system http://www.terason.com.

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004

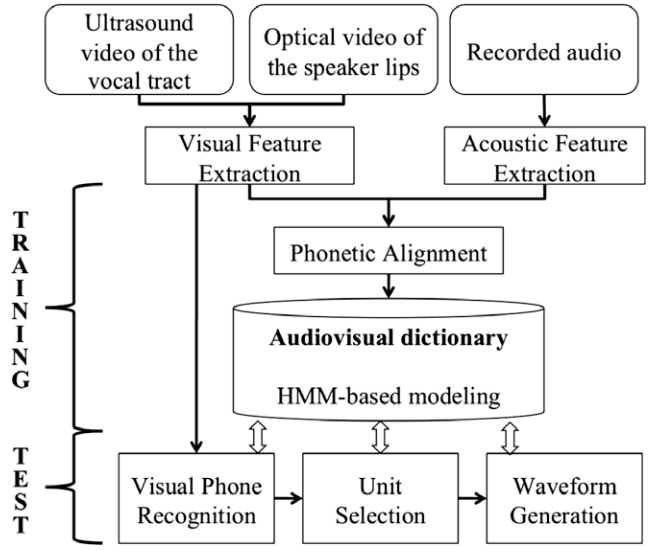

Fig. 2. Overview of the segmental vocoder.

generation procedures, and HNM-based prosody adaptation are presented in the fifth section together with experimental results.

2. Data acquisition

The "Head And Transducer Support" system (HATS, Stone and Davis, 1995) is used for data collection. In this system, the transducer is locked into a fixed position beneath the chin and the head is immobilized. Evolution towards a portable acquisition apparatus is being investigated for a future study. An acoustic standoff is used to minimize or eliminate upward pressure from the transducer on the tissue. Tongue data were collected with an Aloka SSD-1000 ultrasound machine with a multi-frequency (2.5–6 MHz) convex transducer (120° scan angle) which provides a single section of the tongue, in this case the mid-sagittal plane. The ultrasound machine was configured for a single focal point at a 7 cm depth, and post-processing algorithms, such as image averaging and speckle reduction, were disabled. To get both profile and lateral views of the speaker's lips, two standard video cameras were used. A microphone captured the speech signal. The three video streams (two cameras plus ultrasound) and the audio signal are combined together into the same video sequence with an analog video mixer in the NTSC format. This analog acquisition system unfortunately limits the frame rate of the video data to 29.97 Hz since the three video streams become normalized to this rate. A new digital acquisition system, which permits to work at higher frames rates and can handle streams with different frame rates, is currently under study (Ultraspeech system, Hueber et al., 2008a). Typical images recorded by the current, analog acquisition system are shown in Fig. 3.

Because our approach combines phone recognition and diphone-based concatenative synthesis, the textual material of the recorded database must be phonetically balanced and have a good coverage of the diphone space. To meet these two conditions, the CMU-Arctic text corpus, which is used for the design of synthetic voices in the Festvox Text-to-Speech system (Kominek and Black, 2004) was chosen. This corpus contains 1132 sentences divided into two phonetically balanced sets (S1 and S2) containing, respectively, 593 and 539 items. American English is described here by a phoneme set of 41 elements (39 phonemes plus schwa and pause) so that the diphone coverage in sets S1 and S2 is 78% and 75.4%, respectively.

Two native speakers of American English (one female and one male, referred to, respectively, as A and B) were asked to read all sentences of both sets, which were displayed on a laptop screen in front of them. The speakers were instructed to speak as neutrally as possible. Because no inter-session re-calibration mechanism was employed, data was recorded in a single long session during which subjects remained fixed in the HATS system. For both speakers, the full Arctic S1 set was acquired, but because of speaker fatigue, the acquisition of set S2 was limited to 80%, so that the total number of sentences was 1020 rather than the expected 1132.

Due to tissue deformation or abnormal movement of the acoustic standoff, some displacements or drifts may be observed over the course of an acquisition session of several hours. In a head-based coordinate system, potential head and transducer displacements can be detected and

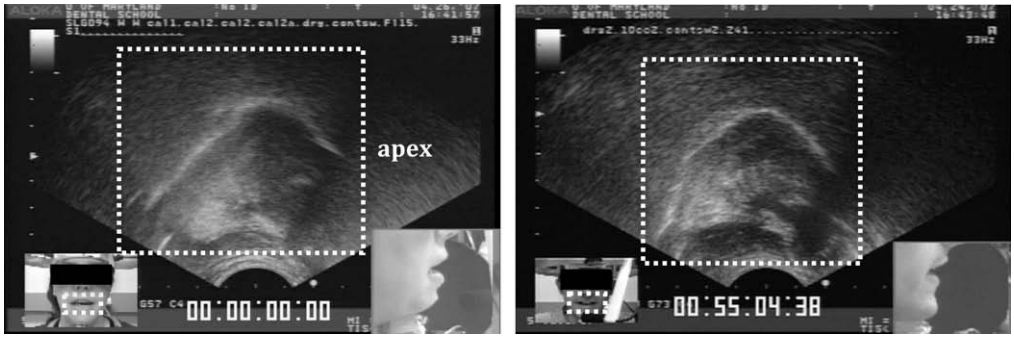

Fig. 3. Example of an ultrasound vocal tract image with embedded lip view and regions of interest for speaker A (left) and speaker B.

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004

monitored by extracting palatal traces at intervals throughout the acquisition. This is because during swallowing, the tongue contacts the roof of the mouth, allowing the ultrasound beam to traverse soft tissue until it is reflected by the palate bone (Epstein et al., 2001). To obtain the palatal traces, 10 cc water deglutitions were executed during brief pauses every 90 sentences. For both speakers, the palatal traces extracted manually from four widely separated deglutitions are super-imposed and shown in Fig. 4.

For speaker A, the proximity of the palatal traces demonstrates that the speaker's head remained stable during the acquisition. However, this is obviously not the case for speaker B, for whom the head/transducer relationship has changed substantially between "Time 1" and "Time 2". This "accidental" displacement, which was noticed and reported by speaker B during the acquisition, occurred after the recording of 240 sentences. From observing the orientation of the hyoid and mandible acoustic shadows in the ultrasound images before and after this displacement, it became apparent that the images corresponding to the first 240 recorded sentences of speaker B were shifted. In order to avoid introducing image degradation with a re-alignment procedure, which would consist here of an image rotation, the first 240 sentences of the speaker B database were simply dropped, and in order to have the same amount of data in both databases, we chose to drop these sentences from the speaker A database as well. Even after removing these data, as compared with speaker A, palatal traces of speaker B between Time 2 and Time 4 still appeared less stable, and periodic alignment problems in fact remained. In a sequence of optical images, where reference points can be tracked (the eyes for instance), such displacements, which are a combination of translations and rotations, can be estimated. However, their accurate estimation in a sequence of ultrasound images, where tracking a tissue point is much more difficult, was considered unfeasible, and in the end the remainder of the speaker B database was left as it was taken. In a general sense, the anatomy of speaker B resulted in his being less well fixed into the acquisition setup, and the problems encountered undoubtedly contributed to the deficit in performance observed for this speaker, as discussed below. There is still much to be learned about the best way to adapt an ultrasound SSI system to each individual user; these concerns will also enter into the design of a future, portable apparatus.

Although it is not a major focus of the present article, we note in passing that stabilizing the position and orientation of the ultrasound transducer and camera with respect to the speaker's vocal tract will definitely be a major challenge. As a first step towards a wearable system, we have been experimenting with 3-axis accelerometers which measure the position of the ultrasound probe with respect of the speaker's head. Such a technique may help re-align the ultrasound images and compensate for relative movement of the probe or optical camera.

After cleaning the data, both speaker databases contained roughly 45 min of speech (780 sentences) and 80,000 bitmap frames. Audio files were sampled at 16,000 Hz.

3. Tongue and lip visual feature extraction

This section describes all steps of the visual feature extraction chain, in which acquired images are processed and coded into vectors of features. First, regions of interest (ROI) are selected in ultrasound and optical images. For ultrasound images, the region of interest is a rectangle delimited horizontally by the acoustic shadows of the mandible and hyoid bone, and vertically by the short tendon, hyoid bone and palate (visible during deglutition), as shown in Fig. 3. If the head position relative to the transducer remains the same during the acquisition, the region of interest determined on the first recorded images is considered valid for the rest of the database. For the lip images, the region of interest is defined as a bounding box including the lips, as shown in Fig. 3. Following the observations of Lucey and Potamianos (2006) and Hueber et al. (2008b), that the frontal view of the lips seems to provide more articulatory information than the profile, only the frontal view will be used in this study, although both are present in the database.

A natural approach for describing an ultrasound image of the tongue is the extraction and the parameterization of the tongue surface contour, for which approaches based on active contours (snakes), often combined with a spline

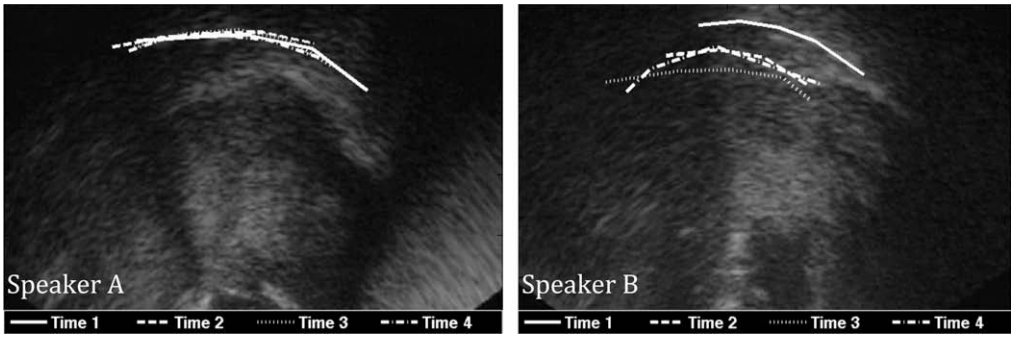

Fig. 4. Palatal traces extracted during four separated deglutitions.

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004

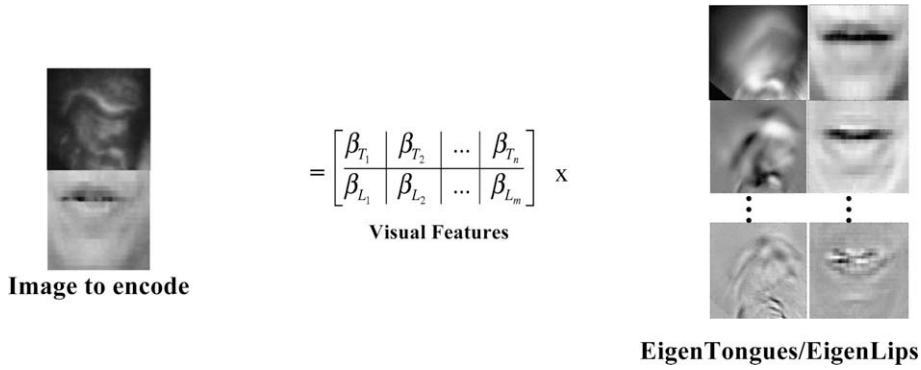

Fig. 5. Encoding ultrasound and optical images of the tongue and lips using the EigenTongue/EigenLips decomposition.

interpolation of the tongue surface, are state of the art (Li et al., 2005; Akgul et al., 2000). However, when nearly parallel to the ultrasound beam, the tongue surface is poorly imaged, as in the case of the phoneme [iy][2] (as in beet) for instance. Contour tracking algorithms are also not efficient enough to cope with small gaps or "double edge" artifacts appearing in the tongue contour, and may fail for such frames. One solution to this problem consists in a more global approach where the entire region of interest is taken into account. In (Hueber et al., 2007a,b), the "EigenTongues" approach, derived from the "EigenFaces" method developed for face recognition (Turk and Pentland, 1991) was introduced. In this technique, each ultrasound image is projected into the representative space of "EigenTongues". This space can be seen as the space of standard vocal tract configurations and is obtained after a Principal Components Analysis (PCA) of a subset of typical frames. In order to guarantee a better exploration of the possible vocal tract configurations, this subset is constructed so as to be phonetically balanced. The EigenTongue components encode the maximum amount of relevant information in the images, mainly tongue position, of course, but also other structures such as the hyoid bone, the position of the short tendon, visible muscles such as the genioglossus and also fat below the tongue. A similar "EigenLips" decomposition is used to encode optical images of the lips. An illustration of the EigenTongue/EigenLips decomposition techniques is given in Fig. 5, where a tongue image and a lip image are represented by their coordinates β_T and β_L.

The authors also tried other dimensionality reduction techniques such as Linear Discriminant Analysis and fast linear transforms (Discrete Cosine Transform). These alternatives however were unable to improve the recognition results.

Before performing the EigenTongue/EigenLips decomposition, ultrasound and optical regions of interest are resized to 64×64 pixel images using cubic interpolation and are converted to grayscale. To reduce the effect of speckle, reduced ultrasound images are often filtered using an anisotropic diffusion filter, which tends to smooth the noise without destroying important features such as edges (Perona and Malik, 1990). The filter proposed by Yu and Acton (2002), was adopted here. The indices (n, m) which quantify the number 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 of the pair (n, m) used on both database are (30, 30). In order to be compatible (albeit artificially) with a more standard frame rate for speech analysis, the sequences of EigenTongues/EigenLips coefficients are oversampled from 30 to 100 Hz using linear interpolation. The effective frame size thus corresponds to 10 ms. Using a "feature fusion strategy", tongue and lip features are concatenated into a single visual feature vector, along with their first and second derivatives, resulting in vectors of typically 180 components. The EigenTongues/EigenLips features are used in the remainder of this article.

4. Visuo-phonetic decoding

4.1. Implementation and evaluation protocol

This section details the visuo-phonetic decoding stage, in which an HMM-based speech recognizer predicts a sequence of phonetic targets from an unknown sequence of visual features. The choice of HMM-based modeling rather than a simple GMM is motivated by the assumption that a visual phone must be interpreted in terms of tongue and lip trajectories, and not only in terms of static articulatory positions. During the training stage, sequences of visual features are modeled, for each phonetic class, by a left-to-right, 5-state (with one non-emitting initial state and one non-emitting terminating state), context independent, continuous HMM. It was assumed that context-dependent models could not have been efficiently trained, even with parameter tying, because of the small size of our training datasets (less than 1 h). Phonetic transcription is derived from the text by using the CMU pronunciation dictionary, which transcribes over 125,000 words into phone sequences. The set of 41 visual phone models is

[2] In this paper, the TIMIT format is used for phonetic transcription.

initialized by a uniform segmentation of the visual data (flat start initialization) and refined using incremental embedded training. During training, the number of Gaussians per state is increased iteratively up to 32 using a dyadic procedure. All HMM work in our study was done using the HTK front-end (Young et al., 2005).

During the testing stage, phonetic decoding is performed using a "Token Passing" algorithm which finds the optimal path through an HMM network (Young et al., 1989). Recognizing accurately a sequence of phones from tongue and lip observations only is a difficult task. Some very important sources of information are missing in the visual data, such as the voiced/unvoiced characteristic and the nasality. Moreover, the tongue visualization is incomplete. The apex (the tongue tip), which is very important in the articulation of dental sounds, is not always correctly imaged in ultrasound imagery, as it is often hidden by the acoustic shadow of the mandible. To overcome these limitations, linguistic constraints can be introduced to help the phonetic decoding. In a traditional "audio speech recognition task", a stochastic language model trained on large text corpora can be used to constrain the HMM decoding. Because the CMU-Arctic corpus was originally designed for speech synthesis, it does not necessarily contain the most likely combination of sub-word units and could even show very rare ones. This is why the use of a universal language model, which *a priori* would not be efficient on that corpus, is not envisioned in this paper.

The size and the content of the dictionary from which the decoding network is built could also have an important impact on system performance. In this study, we propose to use the dictionary as a way of introducing a variable level of linguistic constraint in the HMM decoding process. Three decoding scenarios are introduced:

(a) In the "free phonetic" decoding scenario, the dictionary contains all possible phones; no phonotactic constraint is used; and the grammar from which the decoding network is built allows all possible phoneme combinations.
(b) In the "word-based" scenario, the decoder is forced to recognize words contained in the CMU-Arctic sentences. In this scenario, the word dictionary contains 2569 items and the grammar allows all possible word combinations.
(c) These first two represent the highest and lowest levels of linguistic constraint. In the third, intermediate scenario, called "hybrid" phonetic decoding, the dictionary contains all diphones and triphones observed in the CMU-Arctic corpus, plus all words in the CMU-Arctic sentences, for a total of 12,125 items.

The choice of decoding scenario depends upon the targeted application of the silent vocoder. If it requires a strictly open-domain vocabulary, the system should theoretically be able to decode all possible phoneme combinations, and the first scenario is the most realistic. For a limited-vocabulary application, on the other hand, the second, "strongly constrained" scenario, with its word-based dictionary, will be preferable. Between these two extremes, the third scenario, with its relatively low-level of linguistic constraint, appears most acceptable for a generic application.

For performance evaluation, the 780 sentences of databases A and B are divided into 26 lists of 30 sentences. In order to increase the statistical relevance of the speech recognizer performance, we use a jackknife procedure (Efron, 1981) in which each list is used once as the test set while the other lists compose the training set. For each phone class, a representative measure P of the recognizer performance, the phoneme accuracy, is defined as

$$P = \frac{N - D - S - I}{N}$$

where N is the total number of phones in the reference string of the test set, D the number of deletion errors, S substitution errors, and I insertion errors. According to the Wilson formula (Hogg and Tanis, 1996), a 95% confidence half-interval is computed,

$$\Delta = \frac{t_\alpha \sqrt{P(1-P)/N + t_\alpha^2/(4N^2)}}{1 + t_\alpha^2/N} \quad \text{with} \quad t_\alpha = 1.95$$

and a normal distribution assumption.

Phonetic equivalence classes, in which within-group confusions were not counted as errors, were also created among some of the 41 visual HMMs. These are {[er]-[r]}, {[ih]-[iy]}, {[uh]-[uw]}, {[ao]-[ow]} and {[aa]-[ah]}; an additional class, {[sh]-[zh]}, was added to account for the rare occurrences of the phoneme [zh] (52 times in the database). The jackknife procedure produces 26 (ideally independent) experiments, of which two are kept for validation. The number of Gaussians per state for each recognizer, as well as an inter-HMM transition penalty,[3] used to limit the number of insertions, are estimated using these validation experiments. The authors chose parameters so as to maximize the phoneme accuracy. An alternative approach would be to tune the inter-model transition penalty to balance the insertion and deletion rates. However, in tests using the visuo-phonetic decoding scenarios, several percentage points of recognition accuracy were lost in the process. As discussed below, this suggests that the numerous deletions are "real", and a consequence of the low frame rate of our acquisition system.

In order to compare the visual recognizer performance to a "best possible" result, a traditional phonetic decoder, based on acoustic features derived from the uttered speech signal, is also evaluated. The acoustic component of the audiovisual database was first parameterized using 12 Mel-frequency cepstral coefficients, along with their energies

[3] Typical values for this penalty are −100 (HTK model insertion log probability) for the word-based model and triphone/diphone-based model (hybrid decoding scenario), and −20 for the phone-based model.

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004

and first and second derivatives, and modeled by a set of 41 continuous HMMs (left-to-right, five states, 32 Gaussians per state). An identical procedure as that used in the modeling of the visual features was used here for training, testing and evaluating the performance of this acoustic phonetic decoder.

4.2. Experimental results and interpretations

Table 1 details the performance of the visual phone recognizer, for both speakers and for the three different decoding scenarios.

While it is reassuring to see that the performance is roughly similar for the two speakers, the phone recognizer in fact performs better for speaker A than for speaker B in all decoding scenarios. As seen in Section 2, some experimental issues characterized by small displacements of speaker's head occurred during database B recording. Such shifts can lead to two identical articulatory configurations being imaged differently, thus degrading the coherence of the feature space and in turn, the accuracy of the visual HMMs. To correct for this, both lip images and ultrasound images would need to be re-aligned for this database. Such a procedure is very difficult and, was considered to be too cumbersome for the scope of this article. The interpretations and discussions in the remainder of the article will focus on database A, which was recorded without any of these experimental difficulties.

As expected, the level of linguistic constraints on the HMM decoding has an influence on the system performance. For both audio and video modalities, the best and worst performances are always achieved for the word-based and phone-based decoding scenarios, respectively. Performance obtained in the hybrid decoding scenario, where word and subword-based units are merged in the same dictionary, is always intermediate. This decoding scheme embeds enough linguistic information to constrain the recognition without limiting the available vocabulary, and is thus appropriate for generic applications. For comparison, we also tested a phone-based decoding scenario with a simple bigram language model based on the training set. Using this model, we obtained a 0.7% absolute improvement compared to the "free" phonetic decoding scenario. Such a phonotactic language model does not actually assist the recognizer very much because the CMU corpus was constructed to ideally cover the entire the diphone space, and, as such, it may contain unusual phone sequences. Our hybrid scenario gives better performance.

The performance of the visual phone recognizer is plagued by a large number of deletion errors, especially when decoding is done with a relative low-level of linguistic constraints, as in the phone-based and hybrid scenarios. The phones which are most often deleted are very short ones, corresponding to rapid articulatory gestures such as [t], [d] and [n]. In fact, the phone [t], with a mean duration of 65 ms in our database, is usually represented by only one or two ultrasound images in a 30 fps video sequence, and between 3 and 6 frames after feature interpolation. This phone is thus poorly modeled by a 3-state HMM. This issue is illustrated in Fig. 6, where estimated probability density functions of the five first EigenTongue coefficients, for the three emitting HMM states, are displayed for phones [t] and [ng]. For the phone [ng], with a mean duration of 126 ms in our database, the tongue trajectory is represented by 3 or 4 images (9 to 12 interpolated frames), so that each state of the HMM can model a specific portion of the tongue motion. Thus, the estimated probability density functions differ significantly from one state to another (see Fig. 6), while in the case of the phone [t], they are quite similar. This suggests that we could use smaller HMM models (1 or 2 emitting states) for short phones. We partially circumvent this problem by interpolating the visual features such that the effective frame size corresponds to 10 ms (see Section 3, page 11). In order to reduce the deletion rate, a new acquisition system with higher frame rate (Ultraspeech system, Hueber et al., 2008a) is under study; it is further discussed in the concluding section of the article.

In order to further analyze the quality of the HMM-based modeling, a confusion matrix is derived for the unconstrained phone-based decoding scenario (speaker A), as displayed in Fig. 7.

As expected, the phones which are the most often confused are those with similar articulatory gestures such as {[p],[b],[m]}, {[k],[g],[ng]}, {[f],[v]}, {[s],[z]} and {[t],[d],[n]}. Some of the vowel mismatches are quite "reasonable", such as [uh] (book) being confused with [uw] (boot), and

Table 1
Visual (V) and acoustic (A) based phone recognizer performance.

	Speaker A						Speaker B					
	Phone-based decoding		Hybrid decoding		Word-based decoding		Phone-based decoding		Hybrid decoding		Word-based decoding	
	A	V	A	V	A	V	A	V	A	V	A	V
***P* (%)**	**80.8**	**59.0**	**85.8**	**60.8**	**89.6**	**67.6**	**77.3**	**48.3**	**83.1**	**50.9**	**86.2**	**56.0**
2Δ	1.0	1,2	0.9	1,2	0.8	1,2	1.0	1.3	0.9	1.3	0.9	1.2
D	2087	4614	1780	4711	1305	3383	2397	6362	2145	7040	1550	5388
S	2163	4409	1305	4062	923	3567	2717	5292	1627	4341	1419	4417
I	464	1039	394	846	318	984	459	1027	361	662	406	993
N	24,496											

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004

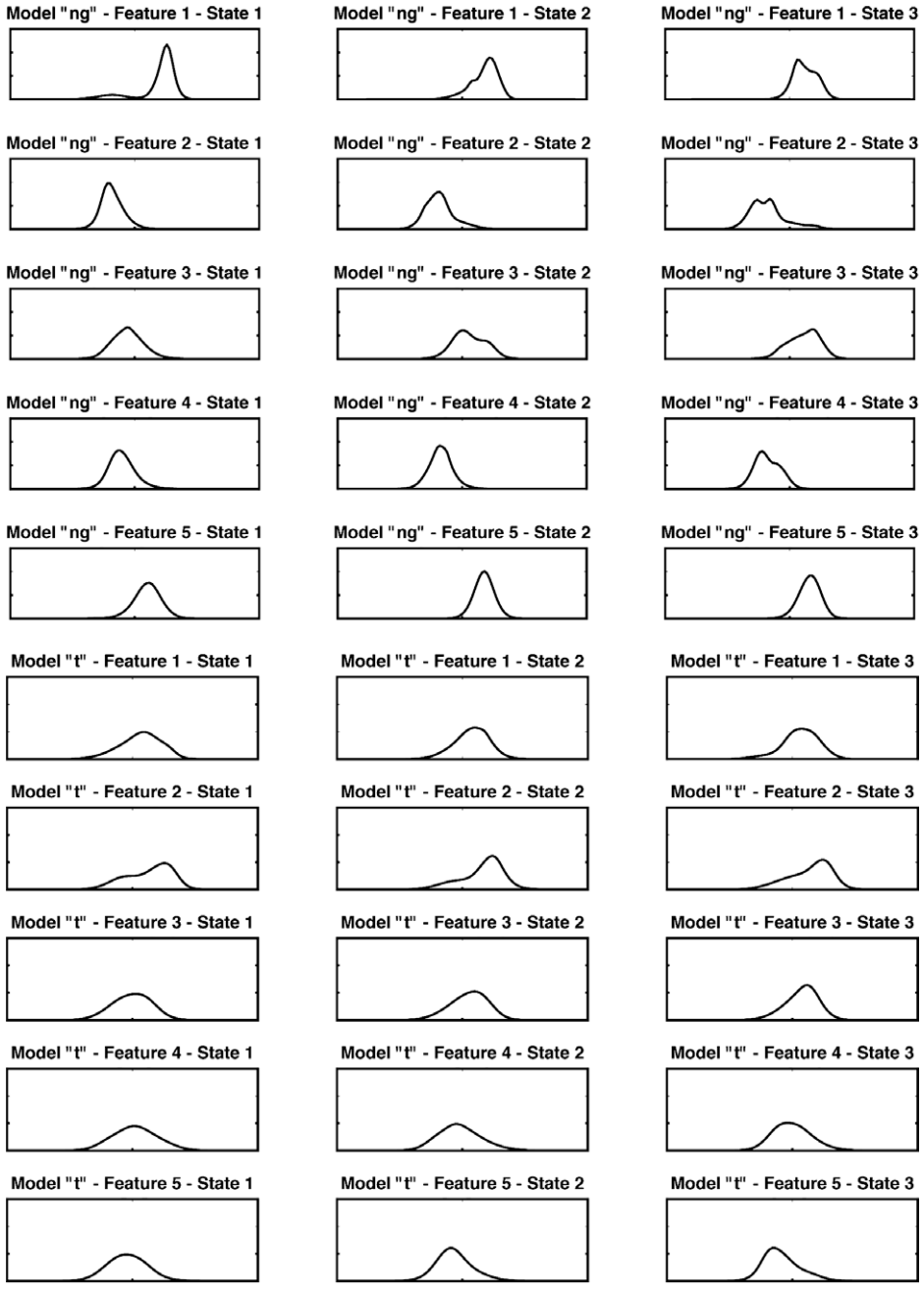

Fig. 6. Estimated probability density function for the five first EigenTongues coefficients (referred as «Feature») of the three emitting HMM states, for the phone [ng] (top) and phone [t] (bottom).

[iy] (beet) interpreted as [ih] (bit). Because of vowel reduction and syllabic consonants (such as the [l] in "bottle"), many phones are confused with the phone [ah]. Diphthongs for which a tongue glide is involved are sometimes confused with one of their pure vowel components, for example [ey] (bait), [oy] (boy) and [ow] may be matched with [ah], [iy] and [ao] (caught), respectively. The matrix also clearly shows an error occurring on dental and alveolar sounds {[th],[dh]} (thin, then) and {[t],[d],[s],[sh]}. This can be explained by the lack of information about the tongue tip (apex) and voicing in the ultrasound images. Most of the errors in the confusion matrix for speaker B (not shown) are of the same type as those discussed for speaker A.

Finally, we study the contributions of video (lips) and ultrasound (tongue) individually to the recognition accuracy, for speaker A and the phone-based scenario, as shown in Table 2.

It appears that decoding using ultrasound (tongue) images alone gives quite a good recognition accuracy, only 3.1% below that of the baseline system. There are two possible explanations for this. First, there seems to be a great deal of redundancy between the two streams, with ultrasound images of the tongue clearly conveying the lion's

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004

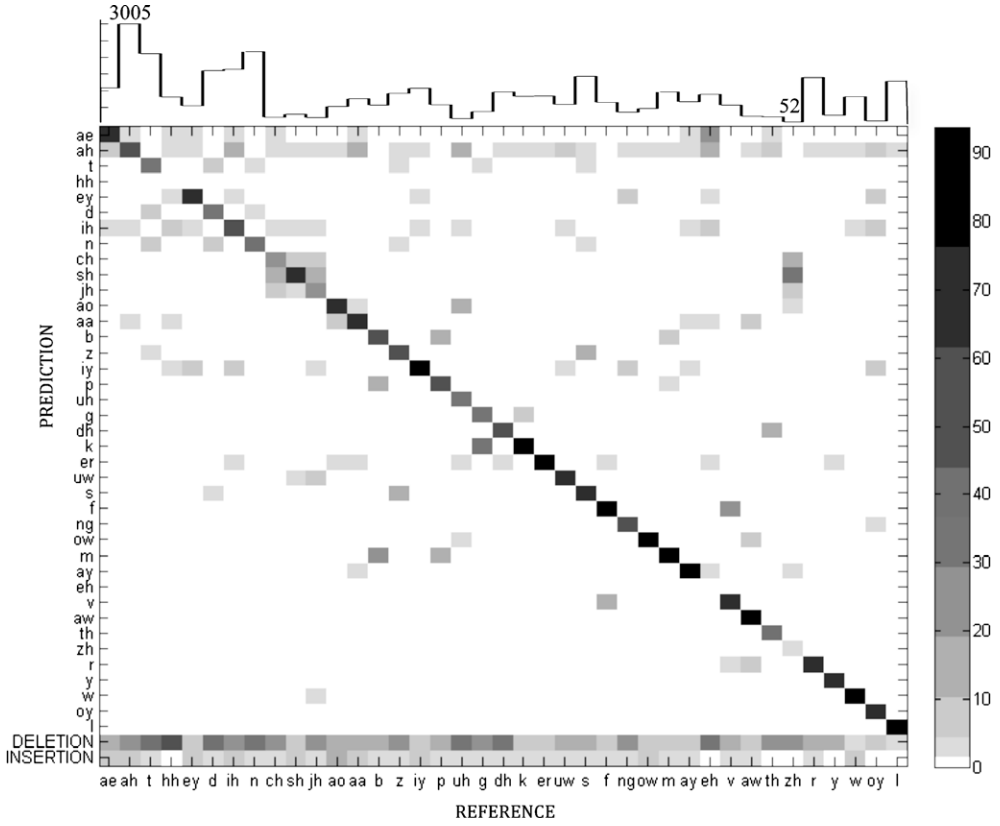

Fig. 7. Confusion matrix for phone recognition from ultrasound tongue sequences and frontal lip views for database A in the phone-based decoding scenario. The color space map was chosen to emphasize errors. The histogram at top shows the number of occurrences of each phone in the database. Most of the errors in the confusion matrix for speaker B (not shown) are of the same type as those discussed for speaker A.

Table 2
Phone recognizer performances using ultrasound (tongue) and video (lips) streams.

	Baseline (tongue + lips)	Ultrasound (tongue)	Video (lips)
***P* (%)**	**59.0%**	**55.9%**	**39.0%**
2Δ	1.2%	1.2%	1.2%
D	4614	5702	6782
S	4409	4277	6805
I	1039	830	1338
N		24,496	

share of the information. Secondly, it is possible that our chosen "feature fusion" technique is not the optimal way of combining the two modalities. We are currently studying more sophisticated fusion techniques such as multi-stream HMMs (Gravier et al., 2002; Hueber et al., 2009).

Though there is still room for improvement, the performance of the visual phone recognizer is promising when compared with the acoustic-based system: for all decoding scenarios, the ratio of the performance of video-based system to that of the audio-based one, which can be interpreted as a target for this database, is about 0.7. Furthermore, some mismatches in the phone recognition stage need not necessarily lead to unintelligible synthesis, and some psychoacoustic effects could potentially also be used to advantage. The results can thus be considered as encouraging enough to warrant investigating the feasibility of a corpus-based synthesis driven by our visual phone decoding system.

5. Corpus-based synthesis driven by video-only data

5.1. Unit selection

In the proposed framework, the segmental vocoder combines phonetic decoding and corpus-based synthesis to recover an acoustic speech signal from video-only data. Our approach is based on the building of an audio–visual dictionary which associates visual units to acoustic units. To initiate the procedure, both video and audio modalities of the recorded database are first labeled at the phonetic level using a forced-alignment procedure. This technique, which is a simplified recognition task where the phonetic sequence is already known, uses both visual and acoustic HMMs estimated during the training of the phonetic decoders. Therefore, the labeled database can be considered as an audiovisual unit dictionary which associates to each visual phone an equivalent in the acoustic domain. As shown in Fig. 8, the temporal boundaries of visual and acoustic phones are not necessarily the same. This asynchrony between the uttered speech signal and the motion of the articulators can be explained by the

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004

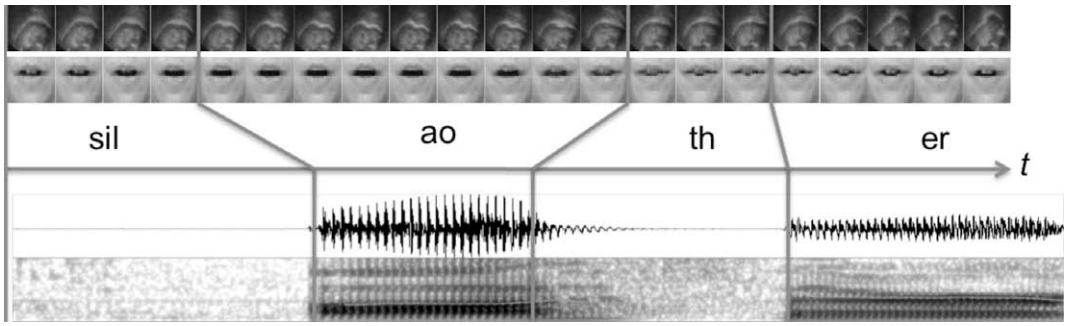

Fig. 8. Asynchronous labeling of the database at the phonetic level.

anticipation phenomena. A segmentation of the dictionary into diphones is then deduced from the phone labeling by requiring spectral stability points at the boundaries of all phones.

Starting from an input sequence of purely visual features, and given the predicted phonetic target, a speech waveform is generated in several steps. A unit selection algorithm first searches for the sequence of diphones that maximizes similarity to the input test data in visual space, while constraining unit concatenation in the acoustic domain. The proposed algorithm, which is an adaptation from the standard path search algorithm used in concatenative speech synthesis (Hunt and Black, 1996), is defined as follows.Assuming a test sequence of visual features $v = [v_1, \ldots, v_L]$, where L is the length of the sequence, and $\tau = [\tau_1, \ldots, \tau_T]$ (with $\tau_1 = 1$ and $\tau_T = L$) the temporal segmentation of v given by the visual phone recognizer, the sequence $T_\tau = \{T_1, \ldots, T_T\}$ of T target units is defined by:

$$T_\tau = \left\{[v_{\tau_1}, \ldots, v_{\tau_2}], \ldots, [v_{\tau_{T-1}}, \ldots, v_{\tau_T}]\right\}$$

Among all appropriate diphone units, the unit selection algorithm finds the optimal sequence $U = \{U_1, \ldots, U_T\}$ that best matches the target T_τ. The quality of the match is determined by two costs, C^t and C^c. The target cost C^t expresses the visual similarity between the target units and the units selected in the dictionary. The target cost between unit U_k and T_k ($1 \leqslant k < T$) is defined by:

$$C^t(U_k,\ T_k) = \mathrm{DTW}(U_k,\ T_k)$$

where $\mathrm{DTW}(a, b)$ is the cumulative distance obtained after a dynamic time warping between the two sequences of visual feature vectors. With this non-linear alignment procedure, temporal stretching and compression observed in the articulator motion are naturally taken into account. The concatenation cost C^c estimates in the acoustic domain the spectral discontinuity introduced when the two units U_k and U_{k+1} are concatenated together and is given by:

$$C^c(U_k,\ U_{k+1}) = D(\mathrm{MFCC}(U_k^{\mathrm{END}}),\ \mathrm{MFCC}(U_{k+1}^1))$$

where D is the Euclidean distance and $\mathrm{MFCC}(U_k^l)$ are MFCC coefficients of the unit U_k at frame l. Because the audio–visual dictionary can be considered as a fully connected state transition network where each state is occupied by a unit, the least costly path that best matches the test sequence can be determined by a Viterbi algorithm (see for instance Forney, 1973). State occupancy is estimated using the visual-based target cost function and transition between states is evaluated by the acoustic-based concatenation cost, as shown in Fig. 9. In the present algorithm, the target and concatenation cost are weighted manually.

5.2. HNM-based speech waveform generation

After the unit selection procedure, the speech waveform can be generated by concatenating the acoustic component of selected diphones. However, no prosodic information such as pitch, loudness or rhythm is available in an SSI, and the generation of an acceptable prosody in a speech synthesis driven only by visual data is a serious issue. Adapting the duration of the selected units can be done according to the temporal segmentation provided by the phonetic decoding. However, there is no guarantee that this "articulatory rhythm" will provide an acceptable "acoustic rhythm". In a syllable-timed language, such as Spanish, French or Japanese, an acceptable solution could be to adapt selected diphone duration so that every syllable of the sentence will take up roughly the same amount of time. However, such a solution cannot be *a priori* adapted to English, which is a stress-timed language (time between two consecutive stressed syllables is more or less constant). Without any information about loudness or pitch, the stressed/unstressed characteristic of the synthesized syllable will potentially have to be inferred from the text. In this study, no duration adaptation mechanism has been implemented, and prosodic modifications are done only in order to generate a speech waveform with an acceptable pitch evolution. Because no pitch information is used during the unit selection stage, the concatenation of selected diphones does not provide a realistic pitch curve. To restore an acceptable prosody, a simple procedure is used. A target sentence having a comparable number of phones, and thus duration, to the test sentence under study (which of course is *not* contained in the training set), is chosen from the training corpus; the pitch pattern of this comparable sentence is then used as a template pattern for the test sentence. The criterion for a sentence to be "comparable" is

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004

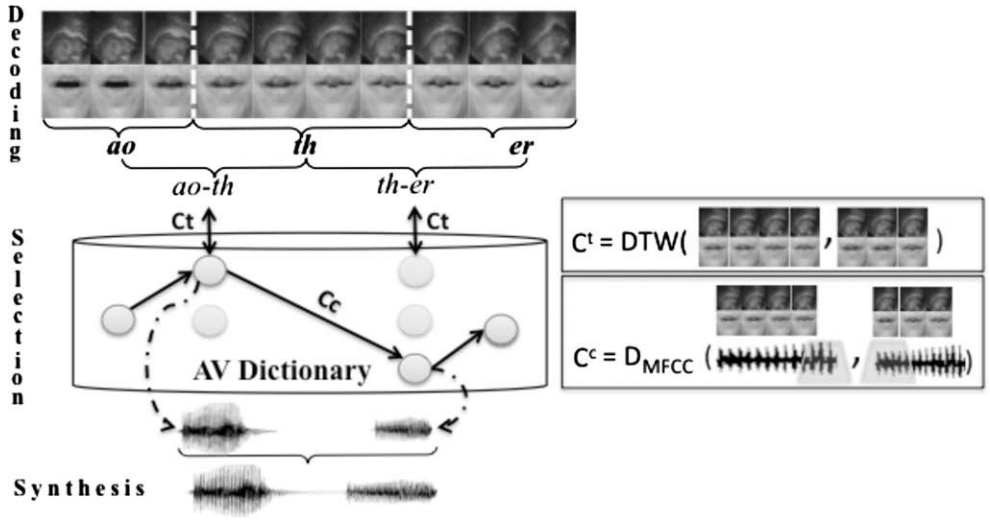

Fig. 9. Phonetic decoding, unit selection algorithm and unit concatenation.

determined simply by the total number of phones, and not the distribution of phones in the sentence. The pitch evolution of the target sentence, hence, is extracted, smoothed using a median filter, and used as a prosody-template for synthesis.

Pitch modification to fit the template, but also loudness and spectrum smoothing at diphone boundaries (the concatenation points), are achieved using a "Harmonic plus Noise Model" of the speech signal (Stylianou et al., 1997). HNM is a pitch-synchronous scheme that allows good-quality prosodic modifications, such as pitch adaptation and time stretching. In the HNM framework, the spectrum of a speech frame $s(t)$ is described as the sum of a harmonic part $H(t)$ and a noise part $B(t)$:

$$s(t) = H(t) + B(t) = \left[\sum_{k=1}^{n} A_t \cos(2\pi k f_0 t + \varphi_k)\right] + \left[B_{\text{gauss}}(t) \times F(t)\right]$$

where n is the number of harmonics included in $H(t)$, f_0 is the estimated fundamental frequency, $B_{\text{gauss}}(t)$ a Gaussian noise frame and $F(t)$ an auto-regressive filter. Our implementation employs 12 harmonic components along with a 16th-order auto-regressive model for the noise part.

5.3. Experimental results

The aim of this subsection is to evaluate our phone-based synthesizer when the phone transcription is 100% correct, independently of the recognition stage, and is in no way meant to be an evaluation of the final performance of our system.

In the proposed framework, the quality of the synthesis obviously depends strongly on the performance of the phonetic decoding stage. In fact, the unit selection synthesis is driven exclusively by the predicted phonetic sequence, and thus an error during the recognition stage will necessarily corrupt the synthesis. In a recent, preliminary study, we have also experimented with HMM-based synthesis (Tokuda et al., 2000), which avoids this "hard decision" problem, and appears thus far to also be promising in our application.

In order to independently evaluate *only* the proposed synthesis framework, i.e., the unit selection algorithm and the HNM-based waveform generation with prosodic adaptation, 20 test sentences are re-synthesized from their visual data alone, after first labeling it with the correct phonetic transcription. The audiovisual unit dictionary is built from the training set, and the 20 test sentences chosen randomly from the test database. Because the phonetic target is given to the unit selection algorithm, this protocol can be seen as a way to evaluate the performance of the vocoder when the phonetic decoder is performing well. Seven native speakers of American English were asked to transcribe the 20 synthesized sentences, as well as the corresponding original sentences. They were allowed to listen to each sentence only once. Clearly in a final system validation, a more realistic intelligibility test, with more listeners, more test data, and unpredictable sentences, should clearly be done, but our test gives a rough idea of what should be possible. While allowing homophones, the quality of the transcription is evaluated with a word-based accuracy criterion, as is traditionally used in speech recognition, and is similar to the criterion P introduced in Section 4.1 (with N now the number of words). The average recognition accuracy of the original sentences was 97%, which gives an idea of the upper bound for this intelligibility test. The results on the synthesized sentences are presented in Fig. 10.

These preliminary results show that the system is able to generate a reasonably intelligible speech waveform from video-only speech data when the phonetic prediction is correct. We also note that short sentences, i.e., with only one prosodic group, are transcribed more accurately, whereas the intelligibility of long sentences is often hindered by a

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004

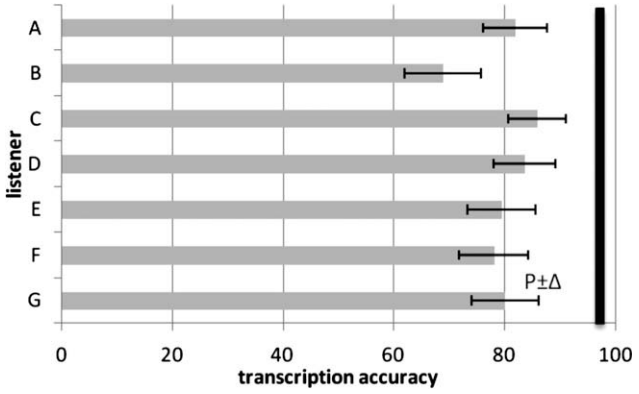

Fig. 10. Evaluation of the synthesized speech intelligibility when the phonetic prediction is correct (transcription accuracy and confidence interval). The black vertical bar indicates the average transcription accuracy of the original sentences (97%).

non-realistic rhythm. With a visual phone recognizer accuracy of 66% though (corresponding to speaker A in the word-based decoding scenario), consistently intelligible synthesis is as yet not possible.

The reader can empirically evaluate the quality, the intelligibility and the naturalness of the synthesis by listening to several examples available at http://www.neurones.espci.fr/ouisper/specom/, where five corpus-based syntheses driven by video-only data are presented. The examples are ordered according to the performance of the visuo-phonetic decoding stage; starting with speech synthesis based on a "100% correct phone recognition", to "typical" performance of the system based on a "60% correct phone recognition". When evaluating these syntheses, it should be remembered that the dictionary is built from less than 1 h of speech data, and thus does not offer the unit selection algorithm a large enough variety of phonetic context to produce a completely natural-sounding result. The acquisition of a larger database will be a critical step. For the moment, though, these examples show that while the system is not yet fully functional, a segmental vocoder driven by visual observations of the tongue and lips is a reasonable approach for making an SSI.

6. Conclusions and perspectives

In the proposed ultrasound-based SSI, a corpus-based speech synthesizer is driven by tongue and lip observations. The segmental vocoder combines an HMM-based visual phone recognition stage with an audiovisual unit selection algorithm and HNM-based prosodic adaptation techniques. In the current framework, the synthesis quality depends strongly on the performance of the visual phone recognizer. As the observation of the tongue and lips alone during speech is probably not enough to consistently achieve accurate phonetic decoding, we have suggested introducing *a priori* linguistic knowledge via the decoding dictionary. The level of linguistic constraint can be easily adjusted, depending on the targeted application. The evaluation of the unit selection algorithm together with HNM-based prosodic adaptation has demonstrated the feasibility of the approach when the phonetic decoder performs well, but with some 60% of phones correctly identified in a sequence of tongue and lip images, the system is not able to systematically provide an intelligible synthesis.

In this paper, the potential of an ultrasound-based SSI has been presented, and its performance evaluated on a difficult recognition task: a continuous speech dataset of only 1 h duration. More experiments still need to be carried out on this data, but better performance could clearly also be obtained on a more limited vocabulary recognition task (less than 250 words, for instance), thus allowing testing of an ultrasound-based SSI in a more restricted, but nevertheless realistic application.

To improve the recognition stage, several solutions are envisioned. First, a new acquisition system has been developed and is currently being tested (Ultraspeech system, Hueber et al., 2008a). Thanks to a higher acquisition rate (above 60 fps for both ultrasound and optical streams), the modeling of very short phones should be more accurate and the number of phone deletion errors reduced. This new system also includes an interactive inter-session re-calibration mechanism, which will allow to record larger audiovisual speech databases in multiple acquisition sessions; the larger unit dictionary thus obtained should improve the performance of both recognition and synthesis stages. The modeling of the tongue and lip motion could also be improved by taking into account possible asynchronies between these two articulators. For that purpose, the use of context-dependent multi-stream HMMs as in (Gravier et al., 2002) is being tested, and preliminary results look interesting (Hueber et al., 2009).

A more straightforward framework, in which the phonetic prediction constrains the dictionary search less, could also be envisioned. To accomplish that, the predicted phonetic target sequence could be enlarged to a lattice of n-best phone sequences, so that the decisions of the unit selection algorithm would rely more on the available units than on the accuracy of the stochastic model predictions. As mentioned earlier, an alternative would be to use HMM-based synthesis (Tokuda et al., 2000) to diminish the problem of the "hard decisions" made in the visuo-phonetic recognition step. We are currently investigating this stochastic synthesis technique, which thus far appears to be promising in our application.

Finally, the system will need to be evaluated on truly "silent" databases, in which sound is either whispered or not vocalized at all. This more realistic experimental protocol will clearly reveal the articulatory specificities of silent speech.

Acknowledgements

This work was supported by the French Department of Defense (DGA), the "Centre de Microélectronique de Paris Ile-de-France" (CEMIP) and the French National

Research Agency (ANR), under the contract number ANR-06-BLAN-0166. The authors would like to thank the anonymous reviewers for numerous valuable suggestions and corrections. They also acknowledge the seven synthesis transcribers for their excellent work, as well as the contributions of the collaboration members and numerous visitors who have attended *Ouisper* Brainstormings over the past 3 years.

References

Akgul, Y.S., Kambhamettu, C., Stone, M., 2000. A Task-specific Contour Tracker for Ultrasound. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, Hilton Head Island, South Carolina, pp. 135–142.

Birkholz, P., Jackèl, D., 2003. A three-dimensional model of the vocal tract for speech synthesis. In: Proc. 15th ICPhS, Barcelona, pp. 2597–2600.

Efron, B., 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika 68, 589–599.

Epstein, M., Stone, M., Pouplier, M., Parthasarathy, V., 2001. Obtaining a palatal trace for ultrasound images. J. Acoust. Soc. Amer. 115 (5), 2631–2632.

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. Med. Eng. Phys. 30 (4), 419–425.

Forney, G.D., 1973. The Viterbi algorithm. Proc. IEEE 61 (3), 268–278.

Gravier, G., Potamianos, G., Neti, C., 2002. Asynchrony modeling for audio–visual speech recognition. In: Proc. 2nd Internat. Conf. on Human Language Technology Research, San Diego, CA.

Heracleous, P., Nakajima, Y., Saruwatari, H., Shikano, K., 2005. A Tissue-conductive Acoustic Sensor Applied in Speech Recognition for Privacy. Smart Objects and Ambient Intelligence Oc-EUSAI, pp. 93–98.

Hogg, R.V., Tanis, E.A., 1996. Probability and Statistical Inference, fifth ed. Prentice Hall, Upper Saddle River, NJ.

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., 2007a. Eigentongue Feature Extraction for an Ultrasound-based Silent Speech Interface. IEEE ICASSP, Honolulu 1, 1245–1248.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2007b. Continuous-speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips. Interspeech, Antwerp, Belgium, pp. 658–661.

Hueber, T., Chollet, G., Denby, B., Stone, M., 2008. Acquisition of Ultrasound, Video and Acoustic Speech Data for a Silent–Speech Interface Application. International Seminar on Speech Production, Strasbourg, France, pp. 365–369.

Hueber, T., Chollet, G., Denby, B., Stone, M., 2008b. Phone Recognition from Ultrasound and Optical Video Sequences for a Silent Speech Interface. Interspeech, Brisbane, Australia, pp. 2032–2035.

Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2009. Visuo-phonetic Decoding using Multi-stream and Context-dependent Models for an Ultrasound-based Silent Speech Interface. Interspeech, Brighton, UK, pp. 640–643.

Hunt, A.J., Black, A.W., 1996. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. IEEE ICASSP, Atlanta, pp. 373–376.

Jorgensen, C., Lee, D.D., Agabon, S., 2003. Sub auditory speech recognition based on EMG/EPG signals. In: Proc. Internat. Joint Conf. Neural Networks, vol. 4, pp. 3128–3133.

Kominek, J., Black, A., 2004. The CMU Arctic speech databases. In: Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, pp. 223–224.

Li, M., Kambhamettu, C., Stone, M., 2005. Automatic contour tracking in ultrasound images. Clin. Linguist. Phonet. 19 (6–7), 545–554.

Lucey, P., Potamianos, G., 2006. Lipreading using profile versus frontal views. In: Proc. IEEE Workshop on Multimedia Signal Processing (MMSP '06), Victoria, BC, Canada, pp. 24–28.

Maeda, S., 1990. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, W., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publishers, Dordrecht, pp. 131–149.

Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., 2005. Session Independent Non-audible Speech Recognition Using Surface Electromyography. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 331–336.

Perona, P., Malik, J., 1990. Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. 12, 629–639.

Sinder, D., Richard, G., Duncan, H., Flanagan, J., Krane, M., Levinson, S., Slimon, S., Davis, D., 1997. Flow visualization in stylized vocal tracts. In: Proc. ASVA97, Tokyo, pp. 439–444.

Stone, M., Davis, E., 1995. A head and transducer support (HATS) system for use in ultrasound imaging of the tongue during speech. J. Acoust. Soc. Amer. 98, 3107–3112.

Stylianou, Y., Dutoit, T., Schroeter, J., 1997. Diphone Concatenation Using a Harmonic Plus Noise Model of Speech. Eurospeech, Rhodes, Greece, pp. 613–616.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech Parameter Generation Algorithms for HMM-based Speech Synthesis. IEEE ICASSP, Istanbul, Turkey, pp. 1315–1318.

Tran, V.-A., Bailly, G., Lœvenbruck, H., Jutten, C., 2008. Improvement to a NAM Captured Whisper-to-Speech System. Interspeech 2008, Brisbane, Australia, pp. 1465–1498.

Turk, M.A., Pentland, A.P., 1991. Face Recognition Using Eigenfaces. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 586–591.

Young, S., Russel, N., Thornton, J., 1989. Token Passing: A Simple Conceptual Model for Connected Speech Recognition Systems, CUED Technical Report F INFENG/TR38, Cambridge University.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2005. The HTK Book. <http://htk.eng.cam.ac.uk/>.

Yu, Y., Acton, S.T., 2002. Speckle reducing anisotropic diffusion. IEEE Trans. Image Proc. 11, 1260–1270.

Please cite this article in press as: Hueber, T. et al., Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips, Speech Comm. (2009), doi:10.1016/j.specom.2009.11.004