OUISPER: CORPUS BASED SYNTHESIS DRIVEN BY ARTICULATORY DATA

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

[1]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
[2]Université Pierre et Marie Curie – Paris VI, 4 place Jussieu, 75252 Paris Cedex 05 France
[3]Laboratoire Traitement et Communication de l'Information, Ecole Nationale Supérieure des Télécommunications (ENST-Paristech), 46 rue Barrault, 75634 Paris Cedex 13 France
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore, MD 21201 USA
hueber@ieee.org, gerard.chollet@tsi.enst.fr, denby@ieee.org, mstone@umaryland.edu, leila.zouari@tsi.enst.fr

ABSTRACT

Certain applications require the production of intelligible speech from articulatory data. This paper outlines a research program (Ouisper : Oral Ultrasound synthetIc SPEech souRce) to synthesize speech from ultrasound acquisition of the tongue movement and video sequences of the lips. Video data is used to search in a multistream corpus associating images of the vocal tract and lips with the audio signal. The search is driven by the recognition of phone units using Hidden Markov Models trained on video sequences. Preliminary results support the feasibility of this approach.

Keywords: clinical phonetics, pathophonetics, speech synthesis, automatic speech recognition

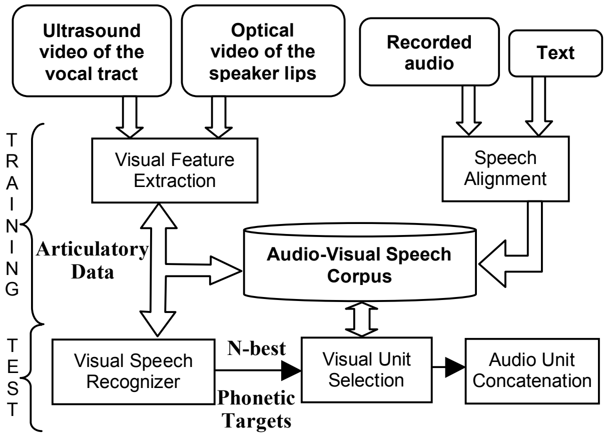
1. INTRODUCTION

The phonemes produced in a language contain a multiplicity of features that are used by the brain to understand and produce speech, but are difficult to reproduce in machine recognition or synthesis. Many articulatory models (CAIP [1], TractSyn [2], Maeda [3]) have focused on rule-based approaches to speech synthesis driven by articulatory parameters. At the same time, the state of the art in text to speech synthesis (for example, the Festival system [4]) uses a corpus-based approach which simply concatenates acoustic speech segments. The Ouisper project proposes to create a speech synthesizer driven by articulatory measurements computed from ultrasound images of the vocal tract and optical images of the speaker lips. It will thus extract discrete phonemes from a continuous data stream and use those as the basis of synthetic speech.

Such a speech synthesizer, driven only by articulatory data, could be used as an alternative to tracheo-oesophageal speech for laryngeal cancer patients, for situations where silence must be maintained, or for voice communication in noisy environments.

Our system is based on the building of an audiovisual corpus which associates articulatory measurements extracted from video to acoustic observations. HMM-based stochastic models, trained on this corpus and combined with a unit selection algorithm, are used to predict and find the optimal sequence of acoustic units, using video-only data. Figure 1 presents an overview of the Ouisper speech synthesis system.

Figure 1: Ouisper corpus-based synthesis system overview



Section 2 of the article details data acquisition and ultrasound image preprocessing, while section 3 describes the visual feature extraction techniques. Speech segmentation is presented in section 4. Visual speech recognition and acoustic unit selection are introduced in section 5.

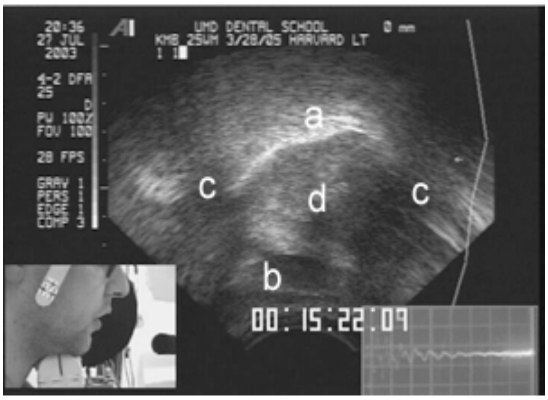
2. DATA ACQUISITION AND PREPROCESSING

The first task of an audiovisual corpus-based speech synthesis system is the construction of an articulatory database comprising video sequences of the voice organ together with the uttered speech signal.

2.1. Data Acquisition

Video sequences of the voice organ are taken using a 30 Hz ultrasound machine and the Vocal Tract Visualization Lab HATS system [5], which maintains acoustic contact between the throat and the ultrasound transducer during speech. A lip profile image is embedded into the ultrasound image, as shown in figure 2.

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



The recorded speech dataset consists of 720 sentences organized in 72 lists from the IEEE/Harvard corpus [6], spoken by a male native American English speaker. The IEEE sentences were chosen because they are constructed to have roughly equal intelligibility across lists and all have approximately the same duration, number of syllables, grammatical structure and intonation. After cleaning the database, the resulting speech was stored as 72473 JPEG frames and 720 WAV audio files sampled at 16000 Hz (43 minutes of speech).

The corpus-based synthesis system currently developed in the Ouisper project provides a general methodology to deal with multimodal corpora. Because this approach is multi-stream based, other data streams can be added, such as dynamic electropalatography (EPG) and electromyography (EMG) [7], or a signal recorded from a "non-audible murmur microphone" [8].

2.2. Ultrasound image preprocessing

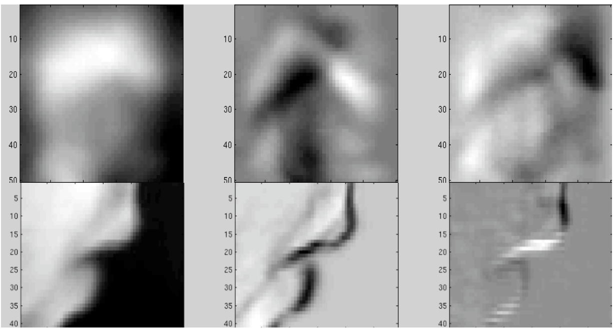
The ultrasound images are first reduced to a polar region of interest grid delimited by the acoustic shadows of the hyoid bone and mandible. The anisotropic diffusion filter of Yu [9] is then applied to remove speckle noise without destroying important image features.

3. VISUAL FEATURE EXTRACTION

3.1. Tongue feature extraction

In many studies (for example [10]), the position of the tongue surface in the image is considered to be the only relevant information in the ultrasound frame, and the extracted articulatory data are simply the parameterized tongue contour. The tongue surface is however poorly imaged when it is nearly parallel to the ultrasound beam, as in the case of the phoneme /i/ for example. Edge tracking algorithms are not enough efficient to cope with small gaps appearing in the tongue contour, and fail for such frames. A solution to this problem is the more global feature extraction approach introduced in [11], wherein Principal Component Analysis (PCA) is used to encode the maximum amount of relevant information in the images, mainly tongue position, of course, but also other structures such as the hyoid bone, muscles, etc. This approach is called "EigenTongues" in analogy to the "EigenFaces" method developed by Turk and Pentland for face recognition [12]. In this way, any vocal tract image is considered to be a linear combination of a set of standard articulatory configurations (*cf.* upper part of figure 3).

Figure 3: The first three EigenTongues (top) / EigenLips (bottom), from left to right.



3.2. Lip feature extraction

To characterize the lip information, a lip contour can of course be used to extract trajectories of the upper/lower lips and commissure from the video sequences. Accurate detection of the lip contour under varying rotations of the speaker face, however, is a difficult task. Hence, a statistical "EigenLip" method was also used to code the lip frames, as illustrated in the lower part of figure 3.

4. SEGMENTAL SPEECH DESCRIPTION

The availability of speech data transcribed at the phonetic level is useful for phonetic research and crucial in the field of corpus based-synthesis. Accurately transcribed speech is needed both for the training of audiovisual speech recognition systems and for the building of a segment database from many transcribed utterances from a single speaker.

4.1. Phonetic segmentation

Manual phonetic segmentation of the speech signal is a difficult and a time consuming task. Several methods have been proposed to speed up this process. The most successful methods have been borrowed from automatic speech recognition, such as Hidden Markov Models (HMM), or Dynamic Time Warping (DTW) techniques, because automatic alignment can be viewed as a simplified recognition task. In this study, an HMM recognizer is used to do forced alignment of speech, that is, a search of the phoneme time boundaries when the phonetic sequence is already known. The speech acoustic signal is parameterized using Mel frequency cepstral decomposition, with normalized energy, delta and acceleration coefficients. HMM acoustic models are initially trained on the transcribed multi-speaker DARPA TIMIT speech database [13].

4.2. Audiovisual database explorer

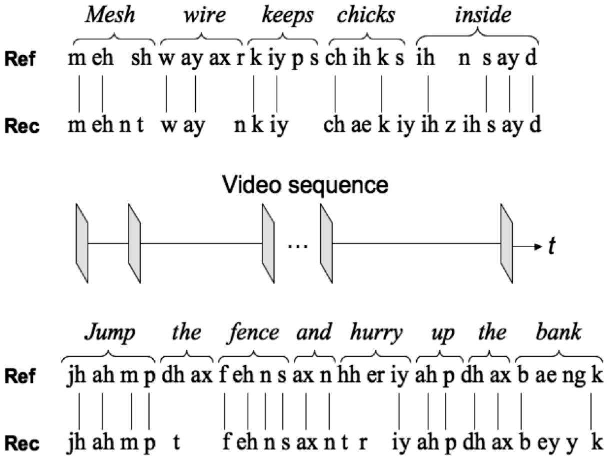
In order to check the speech alignment accuracy and the database coherence, an "audiovisual database explorer" was implemented in the real-time dedicated Max/MSP/Jitter[1] environment. This software allows audiovisual navigation among all of the occurrences of each phoneme classes. For example, a user can listen to all of the /i/ phones, alone or in their context, and simultaneously see the causal motion of the vocal tract and lips.

5. CORPUS BASED SYNTHESIS DRIVEN BY ARTICULATORY DATA

5.1. Speech recognition from video sequences of the vocal tract and lips

During the training phase, visual observation sequences of each phoneme class are modeled using a 5-state, 16 mixture, left-to-right HMM via an embedded re-estimation algorithm. For continuous speech recognition from visual articulatory data, a Viterbi algorithm is used to find the optimal path through the word model network, where word models are obtained by concatenating phone HMM models. In this study, no language model (word sequence probabilities codebook) is used. Thus, this system is driven only by the use of a pronunciation dictionary, which contains in our case 2390 items. Figure 4 illustrates the performance of this visual speech recognizer on two examples, where the predicted phone sequence is time-aligned with the reference phonetic transcription using a dynamic programming-based string alignment procedure.

Figure 4: Reference phonetic transcriptions (Ref) and predicted phonetic transcriptions (Rec) derived from articulatory visual features


Mesh wire keeps chicks inside
Ref m eh sh w ay ax r k iy p s ch ih k s ih n s ay d
Rec m eh n t w ay n k iy ch ae k iy ih z ih s ay d
Video sequence
...
t
Jump the fence and hurry up the bank
Ref jh ah m p dh ax f eh n s ax n hh er iy ah p dh ax b ae ng k
Rec jh ah m p t f eh n s ax n t r iy ah p dh ax b ey y k

Recognition errors are evident in the figure, but our HMM-based system is already able to perform phonetic transcription from *video-only* speech data with over 50% correct recognition. This figure is validated using a jackknife technique, in which each list of ten sentences is used once as the test set. This preliminary result is to be compared to a *best possible* of ≈70% obtained doing traditional

[1] http://www.cycling74.com

speech recognition directly on the *audio* signal, and as such is quite promising.

5.2. Unit selection and concatenation

Given the predicted phonetic sequence, speech synthesis can subsequently be envisioned as selecting phonetic units in the audiovisual corpus. This task is achieved by a Viterbi algorithm, which finds the optimal sequence of visual corpus units that best match the given predicted phonetic sequence. After having selected the visual units, the last step is the concatenation of their correspondents in the acoustical domain. The synthesized speech is of course of good quality for correctly predicted sequences; however, with the current system, the number of errors is still to high to produce a truly usable output signal. One approach will be to enlarge the single target to a lattice of n-best phone sequences, which could be used to drive the search for the optimal solution, a strategy which is an extension of the phonetic vocoder ALISP [14].

6. CONCLUSION AND PERSPECTIVES

The ability to extract discrete phonemes from physiological data is as yet unrealized in speech research. This work is the beginning of such a project and would provide a remarkable enhancement to the current state-of-the-art in speech recognition. The larger goal of the project, to synthesize high quality speech will also be useful to the many applications, commercial and medical, where synthetic speech can augment communication.

Future databases will incorporate a front view of the speaker’s head. Optical flow based techniques will also be used to model the motion of the visible articulators. Speech recognition from video-only data will be validated using a larger dictionary, and may be improved by using a robust language model. Finally, the lack of energy, voicing, and rate information in the video sequence will necessitate the creation of a “virtual prosody” in order to obtain good quality speech synthesis. A data-driven approach, in which prosodic patterns are extracted from the corpus, is foreseen.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge useful discussions with Elsa Angelini and the reviewers for their suggestions. This work is supported by the French Department of Defense (DGA) and the French National Research Agency (ANR).

8. REFERENCES

[1] Sinder, D., Richard, G., Duncan, H., Flanagan, J., Krane, M., Levinson, S., Slimon, S., Davis, D. 1997. Flow Visualization in Stylized Vocal Tracts. *Proc. ASVA97* Tokyo.

[2] Birkholz, P., Jackèl, D. 2003. A three-dimensional model of the vocal tract for speech synthesis. *Proc. 15th ICPhS*, Barcelona, 2597-2600.

[3] 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*, Dordrecht:Kluwer Academic Publishers, 131-149.

[4] Taylor, P., Black, A. and Caley, R. 1998. The architecture of the Festival Speech Synthesis System, *Proc. 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, 147-151.

[5] 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. Am.* 98, 3107-3112.

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

[7] Jorgensen, C., Lee, D. D., Agabon, S., 2003. Sub Auditory Speech Recognition Based on EMG/EPG Signals. *Proc. Int. Joint Conf. on Neural Networks*, vol. 4, 3128-3133.

[8] Nakajima, Y., Heracleous, P., Saruwatari, H., Shikano, K., 2005. A Tissue-conductive Acoustic Sensor Applied in Speech Recognition for Privacy. *Proc Smart Objects and Ambient Intelligences Oc-EUSAI*, 93-98.

[9] Yu, Y., Acton, S. T., 2002. Speckle Reducing Anisotropic Diffusion. *IEEE Trans. on Image Processing*, vol. 11, 1260-1270.

[10] Denby, B., Oussar, Y., Dreyfus, G., Stone, M., 2006. Prospects for a Silent Speech Interface Using Ultrasound Imaging. *IEEE ICASSP*, Toulouse.

[11] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., 2007. Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface. *IEEE ICASSP*, Honolulu.

[12] Turk, M. A., Pentland, A. P., 1991. Face Recognition Using Eigenfaces. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 586-591.

[13] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NTIS order number PB91-100354

[14] Chollet, G., Cernocky, J., Constantinescu, A., Deligne, S., Bimbot, F., 1998. Toward ALISP: Automatic Language Independent Speech Processing. In: Ponting, K., Moore, R. (eds), *NATO-ASI on Speech Pattern Processing*, Berlin: Springer Verlag, 375-388.