# Multimodal HMM-based NAM-to-speech conversion

*Viet-Anh Tran* [1], *Gérard Bailly* [1], *Hélène Lœvenbruck* [1], *Tomoki Toda* [2]

[1] Département Parole & Cognition, GIPSA-lab, UMR 5216 CNRS/INPG/UJF/U. Stendhal
[2] Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{viet-anh.tran, gerard.bailly, helene.loevenbruck}@gipsa-lab.inpg.fr, tomoki@is.naist.jp

## Abstract

Although the segmental intelligibility of converted speech from silent speech using direct signal-to-signal mapping proposed by Toda *et al.* [1] is quite acceptable, listeners have sometimes difficulty in chunking the speech continuum into meaningful words due to incomplete phonetic cues provided by output signals. This paper studies another approach consisting in combining HMM-based statistical speech recognition and synthesis techniques, as well as training on aligned corpora, to convert silent speech to audible voice. By introducing phonological constraints, such systems are expected to improve the phonetic consistency of output signals. Facial movements are used in order to improve the performance of both recognition and synthesis procedures. The results show that including these movements improves the recognition rate by 6.2% and a final improvement of the spectral distortion by 2.7% is observed. The comparison between direct signal-to-signal and phonetic-based mappings is finally commented in this paper.

**Index Terms**: audiovisual voice conversion, non-audible murmur, whispered speech, silent speech interface, HMM-based conversion.

## 1. Introduction

Silent speech consists in articulating sounds with no or little vibration of the vocal cords in order to avoid being overheard [2]. Silent speech is commonly used in situations where private and confidential communication is required. However, it is hard to use it directly in telecommunication, especially with a cellular phone because of its poor intelligibility and unfamiliar perception. This problem challenges researchers with two questions: how to better capture silent speech/ articulation and how to convert it to audible voice? To cope with these challenges, several silent speech interfaces (SSI) have been proposed in the literature: motion capture of fleshpoints on the main speech articulators using Electromagnetic Articulography (EMA) sensors [3], real−time characterization of the vocal tract using ultrasound (US) and optical imaging of the tongue and lips [4][5], digital transformation of signals from a Non Audible Murmur (NAM) microphone [2][1][6][7], surface electromyography (sEMG) of the muscles or the larynx [8][9]. Together with these technologies, two main different approaches have been proposed to generate audible – and visible – speech from signatures of non audible articulation:

1. Plugging a speech synthesis system to a speech recognizer [4][5]. The generation is quite straight-forward: the recognizer segments the speech flow into phonemic units using both signal-dependent information and a more or less sophisticated language model. A standard speech synthesis system then converts this phonetic string into a synthetic voice either using the pre-recorded modal voice of the speaker or built-in available resources. The performance of such a system is mainly dependent on the recognition performance: correct recognition will result in a perfect reconstructed speech while recognition failures or inadequate language models result in drastic degradations.
2. Mapping technique based on GMM model [10][11][1] can be used to directly convert these signals into sound using aligned corpora: joint multi-frame representations of subvocal signals and speech are either stored or modeled and then used to perform direct estimation – or inversion – of speech given the sole representation of subvocal signals. This can be seen as a quantization or optimization process that estimates the most probable speech signal given the subvocal signals and an *a priori* joint model of the combination. The overall quality of the generated speech signals is more homogenous here since the active perception of the listener may compensate for impoverished output signals. No decision is made by the mapping system concerning the phonetic content of the message. Top-down constraints driving speech intelligibility are all provided by the human perceiver.

In both cases, a remaining challenge is the generation of voicing decision and melody – speaker-specific and language-specific tones, accents and intonational patterns – that need to be estimated from non-modal phonation characterized by the absence of vocal folds vibration. Although subvocal articulation seems to still recruit motor neurons driving movements of laryngeal effectors resulting in observable EMG or small displacements of the larynx [12], this “phantom” activity has to be captured and transformed into meaningful melodic movements. So far most systems generate flat melody. Systems combining recognition and synthesis should rely either on language models or recognition of prosodic constituents to drive an intonation model. No such attempts have been reported in the literature so far. Although not completely flat, the synthetic melody computed by voice conversion techniques has a reduced dynamics. First attempts to focus on this generation step have been performed by Tran *et al.* [6]. We notably used large windows over the subvocal signals to estimate suprasegmental features. The naturalness score is noticeably better but there is still much space left for improvement.

In this paper, we focus on the segmental intelligibility of converted speech. We first study the impact of visual information for the HMM-based speech conversion system, for both recognition and synthesis tasks. Then, this system is compared with the GMM-based system proposed by Toda *et al* [1].

The paper is organized as follows. Section 2 describes some characteristics of the NAM microphone. Section 3 describes the HMM-based whisper-to-speech conversion system, the promising contribution of visual information to this system and the comparison between the two approaches mentioned above. Finally, conclusions are drawn in Section 4.

## 2. Non-audible murmur microphone

Nakajima *et al.* [2] proposed a new communication interface which can capture acoustic vibrations in the vocal tract from a sensor placed on the skin, below the ear, called a NAM microphone. This microphone offers a high quality recording of various types of body transmitted speech such as normal speech, whisper and NAM. Body tissue and lip radiation act as a low-pass filter and the high frequency components are attenuated. However, the recorded spectral components still provide sufficient information to distinguish and recognize sound accurately. Currently, the NAM microphone can record sound with frequency components up to 4 kHz. Although this microphone is little sensitive to noise when using simulated noise, its performance decreases in real noise environment because of the Lombard reflex effect [13]. Figure 1 shows an example of whispered speech captured by this microphone. Note that the signal delivered by the NAM microphone is highly sensitive to bursts of stop consonants.

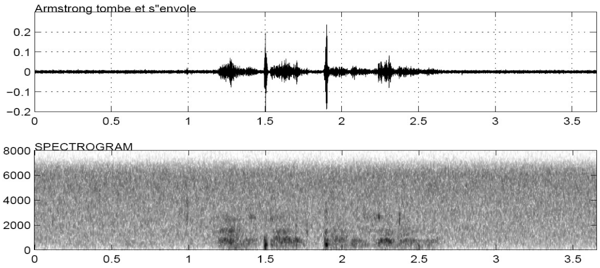


Figure 1: *Whispered speech captured by a NAM sensor for the French utterance: "Armstrong tombe et s'envole" ([amstRõg tõb e sãvol]).*

## 3. Audiovisual HMM-based conversion

During speech production, humans produce sounds by controlling the configuration of oral cavities. The speech articulators determine the resonance characteristics of the vocal tract. Movements of visible articulators such as the jaw and lips are known to significantly contribute to the intelligibility of speech during face-to-face communication. In the field of person-machine communication, visual information can be helpful both as input and output modalities, especially in the case of silent speech [6][7].

### 3.1. Audiovisual corpus

The conversion system is built using audiovisual data pronounced by a native Japanese speaker (the corpus is described in [6]). Two speech modes were recorded: whisper and normal (modal) speech. The system captures, at a sampling rate of 50 Hz, the 3D positions of 142 coloured beads glued on the speaker's face (see Figure 2) in synchrony with the acoustic signal sampled at 16000 Hz.

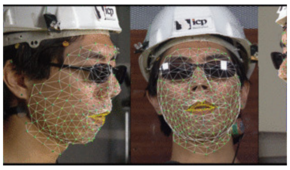


Figure 2*: Characteristic points used for capturing the movements.*

### 3.2. Visual parameters extraction

A shape model is built using a so-called guided Principal Component Analysis (PCA) where *a priori* knowledge is introduced during a linear decomposition. We compute and iteratively subtract predictors using carefully chosen data subsets [14], for a given speaker and a given language. For speech movements and for our particular Japanese speaker, this methodology extracts 5 components that are directly related to the rotation of the jaw, to lip rounding, to upper and lower lip vertical movements and to movements of the throat associated with underlying movements of the larynx and hyoid bone. The resulting articulatory model also includes components for head movements and facial expressions but only components related to speech articulation are considered here.

### 3.3. Conversion system overview

In order to compare the performance of the GMM-based voice conversion technique [1][6] with the approach of combining NAM recognition and speech synthesis, a multi-streams HMM-based whisper-to-speech conversion system was developed. It combines 2 modules, namely HMM recognition and HMM synthesis: instead of the corpus-based synthesis proposed in [5], we use HMM-based synthesis, as described in [15]. The voice conversion is performed in three steps:

1. Using aligned training utterances, the joint probability densities of source and target parameters and duration probability distribution are modeled by context-dependent phone-sized HMM. Static and dynamic acoustic and visual parameters of source and target are stored separately in 4 streams (whispered spectral stream, whispered visual stream, speech spectral stream and speech visual stream). Because of limited training data, we only used the right context for the acoustic models, where subsequent phonemes are classified coarsely into 3 groups for vowels ({/a/}, {/i/,/e/}, {/u/,/o/} without distinguishing between long and short vowels) and 7 groups for consonants: bilabials ({/p/,/pj/},{/b/,/bj/},{/m/,/mj/}), alveolars (/d/,/t/,/n/,/nj/,/s/,/ts/,/z/,/j/), palatals (/ʃ/,/tʃ/,/ʒ/), velars ({/k/,/kj/},{/g/,/gj/}), /f/, /w/ and others ({/h/,/hj/}, {/r/,/rj/}). We add /f/ and /w/ to the context because they are visually distinguished from other consonants (see figure 3). Silences are also classified into 2 groups for utterance-final and internal silences. Gaussian mixtures with two Gaussians and diagonal covariance matrices are used to model the joint observations of each HMM state.

2. HMM-based recognition is performed using the source streams (acoustic and visual) with the HTK toolkit [16]. The linguistic model is limited to phone bi-grams learnt on the training corpus.

3. HMM-based synthesis of the recognized context-dependent phone sequence and target streams (separately acoustic and visual) is performed using the HTS software [15][17].

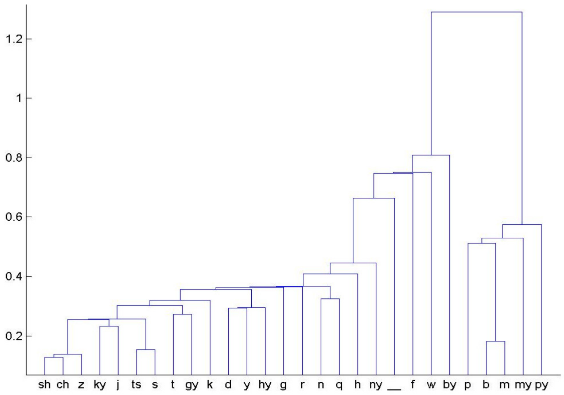


Figure 3: *Confusion tree of whispered visual movements of consonants (the smaller the ordinate, the more confused the two categories are)*

### 3.4. Experiments and results

The Japanese data consists in 150 utterances for training and 40 utterances for the test. The $0^{th}$ through $19^{th}$ mel-cepstral coefficients extracted by STRAIGHT [18] and their first deltas are used as spectral features while 5 visual parameters and their first deltas are used to characterize the movements of the jaw and lips, for both aligned modal speech and whisper.

#### *3.4.1. Impact of visual information for recognition*

Table 1 provides the recognition scores for all phones as well as separately for all vowels and consonants presented in the test corpus. These results show the positive contribution of visual information for the recognition task. On average, all phones considered, the input facial movements improve recognition rate by 6.2 % (65.24 % to 71.43 %). In the case of vowel recognition, the accuracy obtained by using the visual information is 76.45 %, showing an improvement of 8.7 % compared with using acoustic information only. In the case of consonant recognition, this improvement is of 7%. The lesser improvement of consonants compare to that of vowels can be attributed to the large number of labial doubles for Japanese consonants.

Table 2 shows the contribution of facial movements to the recognition of consonants considering the place of articulation. The consonants are classified into 4 groups: bilabials, alveolars, palatals and velars. The bilabials benefit from a very significant improvement (27.6%) while alveolars display only a slight improvement (4.5%). Note that facial movements also benefit surprisingly to the other consonants (17.4% improvement for velars and 14.4% degradation for palatals respectively). The small number occurrences of velars and palatals in the test corpus probably cause this phenomenon. The small facial movements cueing these phones should in fact have no significant impact on their recognition.

Table 1. *Recognition ratio for all vowels, consonants and all the phones represented in the test corpus.*

| Phones | AU (%) | AUVI (%) |
|---|---|---|
| Vowels | 67.79 | 76.45 |
| Consonants | 61.65 | 68.68 |
| All phones | 65.24 | 71.43 |

Table 2. *Recognition ratio with different places of articulation.*

| Phones | AU (%) | AUVI (%) |
|---|---|---|
| **Bilabials** | **53.27** | **80.83** |
| Palatals | 74.98 | 60.6 |
| Alveolars | 67.06 | 71.51 |
| Velars | 63.25 | 80.65 |

Table 3. *Cepstral distortion between converted speech and target speech (dB).*

| System | AU | AUVI |
|---|---|---|
| GMM | 5.99 | 5.77 |
| HMM | 6.58 | 6.4 |

#### *3.4.2. Impact of visual information for synthesis*

The GMM-based system that we used as a reference for this comparison is described in [1][6]. A GMM with 16 gaussians, full covariance matrix is used for the spectral estimation. Global variance is also used to reduce the over-smoothing, which is inevitable in the conventional ML-based parameter estimation [19].

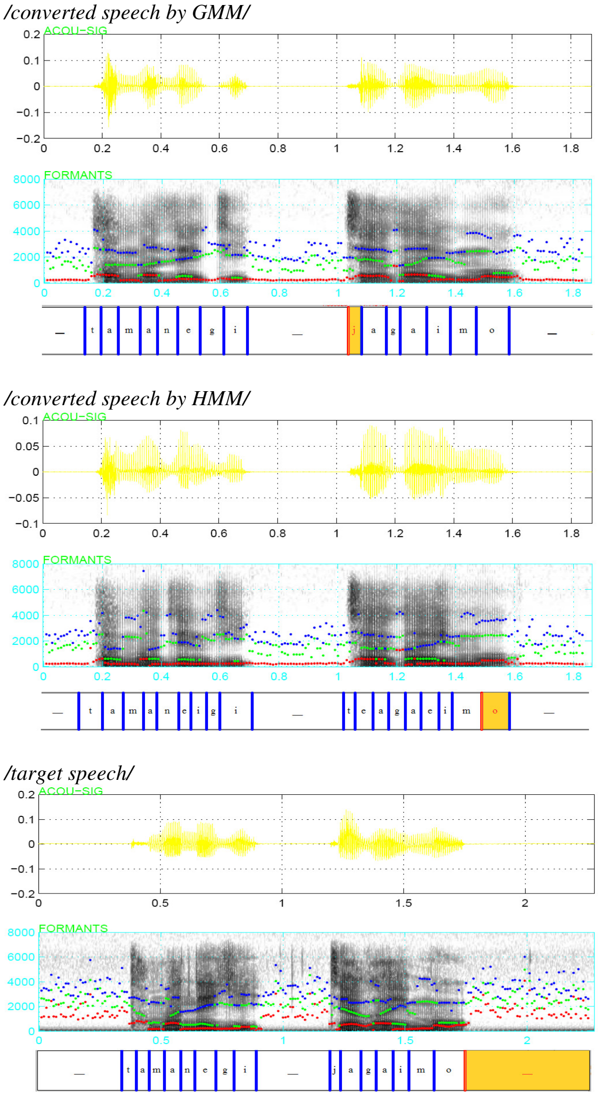
Table 3 compares the contribution of visual information for the intelligibility of converted speech in terms of cepstral distortion between target speech and synthesized speech, with the two systems. Although facial movements have a positive contribution in both systems (cepstral distortion relatively decreases by 2.7% from 6.58 dB to 6.4 dB), the performance of the HMM-based system is currently inferior compared with the direct signal-to-signal system based on GMM model. This inferior score could be explained by two reasons. First, the diagonal covariance currently used for each state of the models in the HMM-based system does not take into account the covariance between whispered speech parameters and speech parameters, but the GMM-based system does, by using a full covariance matrix. Second, synthesis and recognition are used separately, therefore the trained HMM models tend to minimize the recognition error, but not the final reconstruction error.

Figure 4 shows an example of converted speech by the two systems. The formant structures of the GMM-based converted speech is clearer than the other one.

## 4. Conclusions

This paper describes audio-visual whisper to speech conversion that couples a speech synthesis system with a speech recognizer. The facial movements act as a compensation for lip radiation loss in the signal captured by the NAM microphone. This noticeably improves the performance of such a system, especially for the recognition task. The experimental results also show that this influence depends on place of articulation. Although the performance of such a system is currently inferior to the GMM based system, we hope that by modeling the covariance between whispered speech parameters and speech parameters, using more data, extending the acoustic models as well as the linguistic model, and by using global variance, the performance of this system will further improve.

In particular, we think that a more intimate coupling of recognition and synthesis – obtained for example by considering trajectory formation accuracy in HMM training or by considering N-best solutions in the synthesis process – should overcome the limitation of the proposed approach.


*/converted speech by GMM/*

*/converted speech by HMM/*

*/target speech/*

Figure 4: *Whispered speech captured by a NAM sensor for the utterance: "tamanegi jagaimo".*

## 5. Acknowledgements

The authors are grateful to C. Vilain, C. Savariaux, A. Arnal, K. Nakamura for data acquisition, to Prof. Hideki Kawahara of Wakayama University in Japan for the permission to use the STRAIGHT analysis-synthesis system.

## 6. References

[1] Toda, T. and Shikano, K., "NAM-to-Speech Conversion with Gaussian Mixture Models". InterSpeech, Lisbon - Portugal, 1957-1960, 2005.

[2] Nakajima, Y., Kashioka, H., Shikano, K. and Campbell, N. (2003). "Non-audible murmur recognition Input Interface using stethoscopic microphone attached to the skin". International Conference on Acoustics, Speech and Signal Processing, 708-711.

[3] Toda, T., Black A.W., Tokuda K., "Acoustic-to-Articulatory Inversion Mapping with Gaussian Mixture Model", ICSLP, 2004.

[4] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P. and Stone, M., "EigenTongue feature extraction for an ultrasound-based silent speech interface". IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, Hawaii, 1245-1248, 2007.

[5] Hueber, T., Chollet, G., Denby, B., Dreyfus, G. and Stone, M., "Visual phone recognition for an ultrasound-based silent speech interface". Interspeech, Brisbane, Australia, 2008.

[6] Tran, V-A., Bailly, G., Loevenbruck, H. and Toda, T., "Improvement to a NAM captured whisper-to-speech system", Interspeech, 1465-1468, 2008.

[7] Heracleous, P., Beautemps, D., Tran, V.-A., Loevenbruck, H., Bailly, G., "Exploiting visual information for NAM recognition", IEICE Electronics Express, **6**(2): 77-82, 2009.

[8] Jorgensen, C. and Binsted, K., "Web browser control using EMG-based subvocal speech recognition". Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Hawaii, 294c, 2005.

[9] Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F. and Waibel, A., "Towards continuous speech recognition using surface electromyography". InterSpeech, Pittsburgh, PE, 573-576, 2006.

[10] Stylianou, Y., Cappé, O. and Moulines, E., "Continuous probabilistic transform for voice conversion", IEEE Transactions on Speech and Audio Processing, 6(2), 131-142, 1998.

[11] Kain, A. and Macon M.W., "Spectral voice conversion for text-to-speech synthesis", ICASSP, Seattle, 285-288, 1998.

[12] Coleman, J., E. Grabe and Braun, B., "Larynx movements and intonation in whispered speech". Summary of research supported by British Academy (2002).

[13] Heracleous, P., Kaino T., Saruwatari, H., Shikano, K., "Investigating the Role of the Lombard Reflex in Non-Audible Murmur (NAM) Recognition", in Proceedings of Interspeech2005 -EUROSPEECH, pp. 2649–2652, 2005.

[14] Revéret, L., Bailly, G. and Badin, P., "MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation". International Conference on Speech and Language Processing, Beijing, China, 755-758, 2000.

[15] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis". ICASSP,Turkey, 1315–1318, 2000.

[16] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., "The HTK Book". Cambridge, United Kingdom, Entropic Ltd, 1999.

[17] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. and Tokuda, K., "The HMM-based speech synthesis system version 2.0". Speech Synthesis Workshop, Bonn, Germany, 294-299, 2007.

[18] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A. de, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds." Speech Communication 27(3-4): 187-207, 1999.

[19] Toda, T. and Tokuda, K., "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis". Interspeech, Lisbon, Portugal, 2801-2804, 2005.