Predicting F0 and voicing from NAM-captured whispered speech

Viet-Anh Tran, Gérard Bailly*, Hélène Loevenbruck* & Tomoki Toda***

* GIPSA-Lab, Speech & Cognition dpt., 46, av. Félix Viallet, 38031 Grenoble Cedex, France
** 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

The NAM-to-speech conversion proposed by Toda and colleagues which converts Non-Audible Murmur (NAM) to audible speech by statistical mapping trained using aligned corpora is a very promising technique, but its performance is still insufficient, mainly due to the difficulty in estimating F_0 of the transformed voice from unvoiced speech. In this paper, we propose a method to improve F_0 estimation and voicing decision in a NAM-to-speech conversion system based on Gaussian Mixture Models (GMM) applied to whispered speech. Instead of combining voicing decision and F_0 estimation in a single GMM, a simple feed-forward neural network is used to detect voiced segments in the whisper while a GMM estimates a continuous melodic contour based on training voiced segments. The error rate for the voiced/unvoiced decision of the network is 6.8% compared to 9.2% with the original system. Our proposal benefits also to F_0 estimation error.

Keywords: voice conversion, F_0 estimation, neural network, non-audible murmur, whispered speech.

1. Introduction

Speech conveys a wide range of information. Among them, the linguistic content of the message being uttered is of prime importance. However, paralinguistic information such as the speaker's mood, identity or position with respect to what he/she says also plays a crucial part in oral communication [10]. Unfortunately, when a speaker murmurs or whispers, this information is degraded.

To solve this problem, Nakajima *et al.* [5] found that acoustic vibrations in the vocal tract can be captured through the soft tissues of the head with a special acoustic sensor called a NAM microphone attached to the surface of the skin, below the ear. Using this stethoscopic microphone to capture non-audible murmur, Toda *et al.* [1] proposed a NAM-to-Speech conversion system based on the GMM model in order to convert "non-audible speech" to ordinary speech. It was shown that this system effectively works but its performance is still insufficient, especially in the naturalness of the converted speech. This is due to the difficulties in F_0 estimation from unvoiced speech. These authors claimed that it is inevitable to improve the performance of NAM-to-Speech systems. Nakagiri *et al.* [4] proposed another system which converts NAM to whisper. F_0 values do not need to be estimated for converted whispered speech because whisper is another type of unvoiced speech, just like NAM, but more intelligible.

Another direction of research consists in using a phonetic pivot by combining speech recognition and synthesis techniques as in the Ouisper project [12]. By introducing higher linguistic levels, such systems can potentially predict a phonological structure that can be used in speech resynthesis. But no results have been reported yet and such an approach seems unsuitable for applications with open domain.

In this paper, we propose to improve signal-based GMM mapping by a better estimation of the voiced source of the converted speech. Whisper-to-speech was used because of difficulties in getting accurate phonetic segmentation in NAM.

In the training stage, whispered speech and ordinary speech utterance pairs were carefully aligned using phonetic transcription information. The main difference between our system and the original system proposed in [1] is that only voiced segments were provided as input to train the GMM model which maps spectral vectors of whisper to F_0 values of converted speech. In the conversion stage, we use a feed-forward neural network to detect the voiced segments in whispered utterance and then compute F_0 for these segments only instead of computing these values for all segments.

Another innovative aspect of this paper is the language: we have applied voice conversion techniques to French, which has much more complex syllabic structures and a larger phonemic inventory than Japanese. Degraded performance is thus expected with respect to original published results.

The paper is organized as follows. Section 2 describes some characteristics of whispered speech. Section 3 describes the frameworks of our NAM-to-Speech conversion system. Section 4 describes our experimental evaluations and finally, conclusions are drawn in Section 5.

2. Whispered speech

In recent years, advances in wireless communication technology have led to the widespread use of mobile phones for private communication as well as information access using speech. Speaking loudly to a mobile phone in public places may be a nuisance to others, however, whispered speech can only be heard by a limited set of listeners surrounding the speaker and can therefore effectively be used for quiet and private communication over mobile phones [7].

2.1. Acoustic features

In normal speech, voiced sounds involve a modulation of the air flow from the lungs by vibrations of the vocal folds. However, there is no vibration of the vocal folds in the production of whispered speech. Exhalation of air is used as the sound source, and the shape of the pharynx is adjusted such that the vocal folds do not vibrate. Due to this difference in production mechanism, the acoustic characteristics of whisper differ from those of normal speech. A study on the acoustic properties of vowel sounds [7] has shown an upward shift of the formant frequencies for vowels in whispered speech compared to normal speech. The shift is larger for vowels with low formant frequencies. The authors also found that the cepstral distances between normal and whispered speech for vowels and voiced consonants are higher than those of unvoiced consonants. This means that the vocal tract characteristics of vowels and voiced consonants change more significantly in whisper relative to ordinary speech than those of unvoiced consonants.

The perception of vowel pitch in normal speech is related mainly to the fundamental frequency (F_0) which corresponds to periodic pulsing. In whispered speech, however, although

there is no periodic pulsing, some pitch-like perception may occur. Higashikawa *et al.* [8] have shown that listeners can perceive pitch during whispering and formant frequency could be one of the cues used in perception. More precisely, the authors in [9] indicated that "whisper pitch" is more influenced by simultaneous changes in F1 and F2 than by changes in only one of the formants.

2.2. NAM microphone

Nakajima *et al.* [5] proposed a new communication interface which can capture acoustic vibrations in the vocal tract from a sensor placed on the skin, below the ear. This position is shown in Fig. 1. This position allows a high quality recording of various types of body transmitted speech such as normal speech and whisper. Body tissue and lip radiation act as a low-pass filter and the high frequency components are attenuated. However, the non-audible murmur spectral components still provide sufficient information to distinguish and recognize sound accurately [6]. Currently, the NAM microphone can record sound with frequency components up to 4 kHz while being little sensitive to external noise.

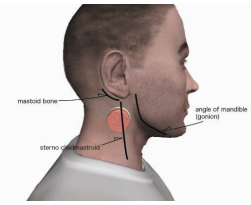


Figure 1: *Position of NAM microphone*.

3. NAM-to-Speech conversion system

Several approaches have been proposed to convert non audible speech to modal voice. The most trivial system consists in chaining NAM recognition with speech synthesis. Direct signal-to-signal mapping using aligned corpora is also very promising: Toda *et al.* [1] applied statistical feature mapping [10][11] to NAM-to-speech conversion.

Although the segmental intelligibility of synthetic signals computed by statistical feature mapping is quite acceptable, listeners have difficulty in chunking the speech continuum into meaningful words. A large part of this problem is due to impoverished synthetic intonation. In this study, we focus on improving the estimation of pitch and voicing of the converted speech. Fig. 2 shows the conversion used in our system.

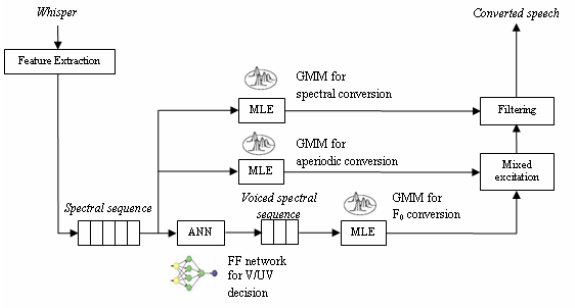


Figure 2*: Conversion process of NAM-to-Speech system. Spectral Estimation.*

Before training the models for spectral estimation and F_0 estimation, the pairs of whisper and speech uttered by a speaker must be aligned because of different speaking rates. Transcription information was used for this task to get a better alignment compared to blind dynamic time warping (DTW).

3.1. Spectral Estimation

We use the same schema for spectral estimation as the one proposed by Toda [1]. As described in [1], feature vector X_t of whisper consisting of spectral feature vectors of several frames around a current frame t was aligned with a target speech feature $Y_t = [y_t, \Delta(y_t)]$ consisting of static and dynamic features. These vectors were then used to train a GMM for representing the joint probability density $p(X_t, Y_t \mid \Theta)$, where Θ denotes a set of GMM parameters. Another model for representing the probability density of global variance (GV) of the target static features $p(v(y) \mid \Theta_v)$ was trained where Θ_v denotes a set of parameters of a Gaussian distribution, $v(y)$ denotes global variance over the time sequence y of target static feature. This global variance information is used to alleviate the over-smoothing, which is inevitable in the conventional ML-based parameter estimation [2].

In the conversion, the target static feature y was estimated from source feature $X = [X_1, X_2, \ldots, X_T]$ so that a likelihood $L = p(Y|X, \Theta)^w p(v(y) \mid \Theta_v)$ was maximize where w is a weight and the vector Y is represented as $W\,y$, where W denotes a conversion matrix from the static feature sequence to the static and dynamic feature sequence.

3.2 Excitation Estimation

The mixed excitation is defined as the frequency-dependent weighted sum of white noise and a pulse train with phase manipulation. The weight is determined based on an aperiodic component in each frequency band [14].

Aperiodic estimation was done in the same way as the spectral estimation except that global variance (GV) was not used because GV does not cause any large difference to the converted speech in the aperiodic conversion [3].

ML-based conversion method was used the F_0 estimation. Static and dynamic features Y_t of F_0 are used while keeping the same feature vector of whisper X_t as that used for the spectral conversion. However, instead of using all the segments in each pair of utterance, only voiced segments X_t, Y_t were extracted to train a GMM on the joint probability in a similar way as the spectral estimation in order to avoid loosing some Gaussian components for representing the zero values of F_0 set for unvoiced segments. A feed-forward neural network is used to predict these segments from X. For synthesis, continuous F_0 values are predicted that are paced by the voicing parameter computed by the network.

4. Evaluation

In order to show our improvement in F_0 estimation and voicing attribution, two evaluations were done, comparing our system with the original system proposed in [1].

The training corpus consists in 200 utterance pairs of whisper and speech uttered by a French male speaker and captured by a NAM microphone and head-set microphone. Respective speech durations are 4.9 minutes for whisper (9.7 minutes with silences) and 4.8 minutes for speech (7.2 minutes with silences). The 0^{th} through 24^{th} mel-cepstral coefficients were used as a spectral feature at each frame. The spectral segment feature of whisper was constructed by concatenating feature vectors at current ± 8 frames, and then the vector dimension was reduced to 50 using a PCA technique. Log-scaled F_0 extracted by STRAIGHT [13] was used as the target feature.

4.1. Voicing estimation

In the original system [1], first, the authors took all the frames in each utterance and estimated F_0 value at each frame by using the trained GMM model. Then, they used a threshold to assign the voiced/unvoiced label for this frame. The F_0 values in every unvoiced frames were then set to zero. We applied this original technique to our data.

Then to compare with our technique, we created a feed-forward neural network with 50 input nodes, 17 hidden nodes and 1 output node. The segmental features at each frame of the whispered utterances were used as input vector for this network. The voiced/unvoiced label for each segment in the training whispered data was obtained from the voiced/ unvoiced label of the corresponding speech utterance by aligning the two utterances. All the whispered utterances used for training the GMM were also used to train this network.

Table 1: *Voicing error using neural network or GMM.*

Type of error	Feed-fwd NN (%)	GMM (%)
Voiced error	2.4	3.3
Unvoiced error	4.4	5.9
Total	~ 6.8	~ 9.2

Table 1 shows the evaluation of this network. Compared with the error in the original system, the result shows that we have a slight improvement of the voiced/unvoiced detection.

4.2. Parameter evaluation

We also compared the two systems with different number of mixtures for estimating F_0 on both the training and the test data. The number of mixtures of mel-cepstral mapping function was set to 32. Full covariance matrices were used for both GMMs. The test corpus consisted of 70 utterance pairs not included in the training data. The error was calculated as the normalized difference between synthetic F_0 and natural F_0 in the voiced segments that were well detected by the two systems. The error is given by the following formula: $Err = (synthetic\_F_0 - natural\_F_0)/ natural\_F_0$. Fig. 3 shows that the proposed framework outperforms the original system. In addition, when the number of Gaussian mixtures increase, the errors of both systems on the training data decrease, but these errors on test data are little sensitive to the number of mixtures. This is further illustrated in Table 2 which provide correlation coefficients for the two systems.

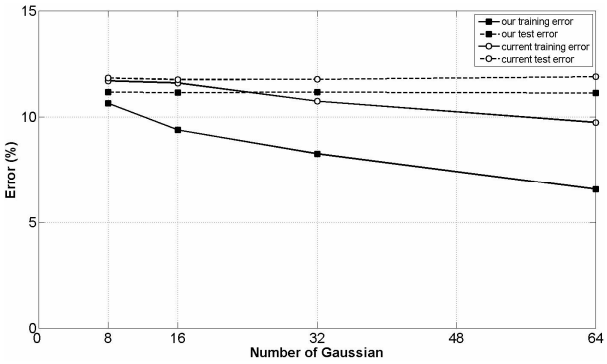


Figure 3: *Natural and synthetic F_0 curve.*

Fig. 4 shows an example of whispered-, converted- (our system) and natural-speech. As can be seen the formant patterns of converted speech are flatter than those of natural speech. Global variance was used to attenuate this difference.

Table 2: *Correlation coefficient between natural F_0 and converted F_0 by the two systems.*

# of Gaussian mixtures	Our system		Original system	
	train	test	train	test
8	0.565	0.495	0.494	0.451
16	0.667	0.493	0.513	0.456
32	0.746	0.498	0.603	0.460
64	0.834	0.499	0.682	0.444

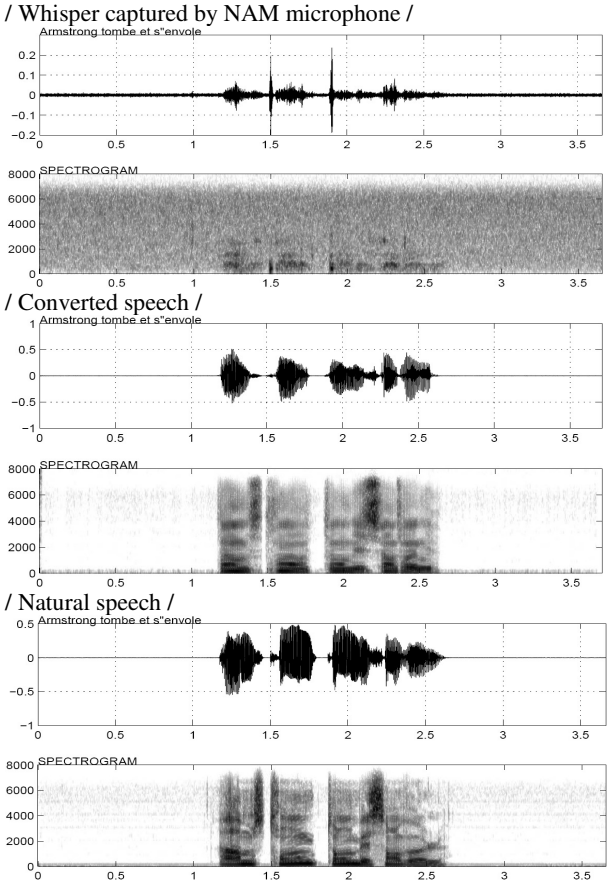


Figure 4: *Whispered speech captured by NAM sensor, converted speech and ordinary speech for the same utterance: "Armstrong tombe et s'envole".*

Figure 5 shows an example of a natural (target) F_0 curve and the synthetic F_0 curves generated by the two systems. It shows that our new system is closer to the natural F_0 curve than the original system.

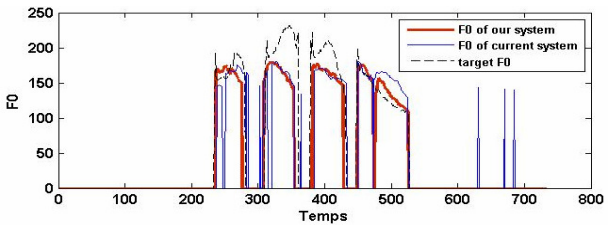
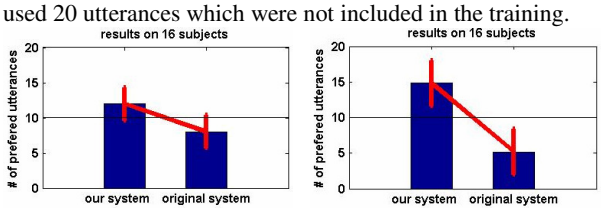


Figure 5: *Natural and synthetic F_0 curve for the same utterance : "Armstrong tombe et s'envole".*

4.3. Perceptual evaluation

Sixteen French listeners who had never listened to NAM participated in our perceptual tests on intelligibility and naturalness of the converted speech from the two systems. We used 20 utterances which were not included in the training.



(a) *(b)*

Figure 6*: Intelligibility score (a) and Naturalness score (b).*

4.3.1. Evaluation on intelligibility

For this evaluation, we used synthetic speech with the estimated mel-cepstrum, estimated aperiodic component and estimated F_0 by both systems.

Each listener heard an utterance pronounced in modal speech and the converted utterances obtained from the whispered speech with both systems. For each utterance, they were asked which one was closer to the original one, in terms of intelligibility (ABX test). Fig. 6a provides the mean intelligibility scores for all the listeners for the converted sentences using the original system and our new system. The intelligibility score is higher for the sentences obtained with our new system ($F = 23.41$, $p < .001$).

In the evaluation in [1], it was shown that F_0 estimation (compared with using a constant F_0) improves intelligibility. This might be caused by a better detection of the voiced/ unvoiced property in the F_0 estimation. In our case, a feed-forward neural network was used instead of using GMM.

4.3.2. Evaluation on naturalness

We used here the version of the test utterances produced with the modal voice. These utterances are considered as ideal desired targets of the mapping. We thus further processed the synthetic utterances by warping their time scale to the targets using the warping procedure used for training.

We then conducted an ABX test. Because of the warping, all utterances have almost the same temporal organization. For each sentence, subjects choose A or B as the nearest to the natural X in terms of naturalness. Fig. 6b shows the mean naturalness that all the listeners rate as the nearest. Again the proposed system was strongly preferred to the original one ($F = 74.89$, $p < .001$).

5. Conclusions

This paper described the improvement in F_0 estimation and voicing decision we propose for NAM-to-speech conversion system applied to whispered speech. GMM models were used to estimate the spectra, aperiodic component and F_0 of the converted speech from spectral segments obtained from NAM-captured whispered speech, based on a maximum likelihood criterion. To estimate the F_0 features in the whispered utterances, only voiced segments were used. They were detected using a simple feed-forward neural network. Although the performance of the system is improved compared to that of the original system, the estimated pitch is still flat due to the GMMs. In the future, we will investigate how to obtain audible speech from whisper by using a HMM which is appropriate for modelling a time sequence of speech parameters. Also, we plan to use complementary information such as video in the aim of obtaining other useful parameters.

Acknowledgment: The authors are grateful to Coriandre Vilain and Alain Arnal for data acquisition, Prof. Hideki Kawahara of Wakayama University in Japan for the permission to use the STRAIGHT system.

6. References

[1] Toda, T.; Shikano, K., 2005. NAM-to-Speech Conversion with Gaussian Mixture Models. In *Proc. Interspeech*. Lisboa, 1957-1960.

[2] Toda, T.; Black, A.W.; Tokuda, K., 2007. Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory. In *IEEE Transactions on Audio, Speech and Language Processing*. Vol. 15, No. 8, 2222-2235.

[3] Ohtani , Y.; Toda, T.; Sarawatari, H.; Shikano, K., 2006. Maximum Likelihood Voice Conversion Based on GMM with STRAIGHT Mixed Excitation. In *Proc. Interspeech – ICSLP*. Pittsburgh, USA. 2266-2269.

[4] Nakagiri, M. ; Toda, T. ; Kashioka, H.; Shikano, K., 2006. Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion. In *Proc. Interspeech-ICSLP*. Pittsburgh, USA. 2270-2273.

[5] Nakajima, Y.; Kashioka, H.; Shikano, K.; Campbell N., 2003. Non-audible murmur recognition. In *Proc. Inter-speech(Eurospeech)*. Geneva, Switzeland, 2601-2604.

[6] Heracleous, P.; Nakajima, Y., 2004. Audible (normal) speech and inaudible murmur recognition using NAM microphone. In *EUSIPCO*, Vienna, Austria.

[7] Ito, T.; Takeda, K.; Itakura, F., 2005. Analysis and recognition of whispered speech. In *Speech Communication*. Lisboa. Vol. 45, Issue 2, 139-152.

[8] Higashikawa, M.; Nakai, K.; Sakakura, A; Takahashi, H., 1996. Perceived Pitch of Whispered Vowels – Relationship with formant frequencies: A preliminary study. *Journal of Voice*, 155-158.

[9] Higashikawa, M; Minifie, F .D., 1999. Acoustical-perceptual correlates of "whispered pitch" in synthetically generated vowels. In *Journal of Speech, Language, and Hearing Research*. Vol 42, 583-591.

[10] Stylianou, Y. ; Cappé O. ; Moulines E, 1998. Continuous probabilistic transform for voice conversion. In *IEEE Trans. Speech and Audio Processing*, Vol. 6, No.2, 131-142.

[11] Kain, A.; Macon M. W., Spectral voice conversion for text-to-speech synthesis. In *Proc. ICASSP*. Seattle, U.S.A. Vol 1, 285-288.

[12] Hueber, T.; Chollet, G.; Denby, B.; Dreyfus G.; Stone M, 2007. Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips. In *Proc. Interspeech*. Antwerp, Belgium,

[13] Kawahara, H.; Masuda-Katsuse, I.; Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time frequency smoothing and instantaneous-frequency- based F_0 extraction: Possible role of a repetitive structure in sounds. In *Speech Communication*. Vol. 27, No. 3-4, 187-207.

[14] Kawahara, H.; Estill, J.; Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. *MAVEBA,* Firentze, Italy.