Improvement to a NAM captured whisper-to-speech system

Viet-Anh Tran [1], *Gérard Bailly* [1] *Hélène Lœvenbruck* [1] *Christian Jutten* [2]

1. Département Parole & Cognition, GIPSA-lab, UMR 5216 CNRS/INPG/UJF/U. Stendhal
2. Département Image & Signal, GIPSA-lab, UMR 5216 CNRS/INPG/UJF/U. Stendhal

{viet-anh.tran, gerard.bailly, helene.loevenbruck, christian.jutten}@gipsa-lab.inpg.fr

Abstract

In this paper, new techniques to improve whisper-to-speech conversion are investigated, in the framework of silent speech telephone communication. A preliminary conversion method from Non-Audible Murmur (NAM) to modal speech, based on statistical mapping trained using aligned corpora has been proposed. Although it is a very promising technique, its performance is still insufficient due to the difficulties in estimating F_0 from unvoiced speech. In this paper, two distinct modifications are proposed, in order to improve the naturalness of the synthesized speech. In the first modification, LDA (Linear Discriminant Analysis) is used instead of PCA (Principal Component Analysis) to reduce the dimensionality of the input spectral vectors. In addition, the influence of long-term variation of spectral information on pitch estimation is examined. The second modification is an attempt to integrate visual information as a complementary input to improve spectral estimation, F_0 estimation and voicing decision.

Index Terms: audiovisual voice conversion, 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 [12]. Unfortunately, when a speaker murmurs or whispers, this information is degraded.

To solve this problem, Nakajima *et al.* [10] 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 Non-Audible Murmur (NAM) microphone attached to the surface of the skin. Using this stethoscopic microphone to capture non-audible murmur, Toda *et al.* [14] proposed a NAM-to-Speech conversion system based on the GMM model in order to convert "non-audible speech" to modal speech. It was shown that this system effectively works but the naturalness of the converted speech is still unsatisfactory. This is due to the poor F_0 estimation from unvoiced speech. These authors conclude that it is necessary to improve the performance of NAM-to-Speech systems. Nakagiri *et al.* [9] propose to simply convert 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 [6]. By introducing higher linguistic levels, such systems can potentially predict a phonological structure that can be used in speech resynthesis. Although excellent results have been reported for Japanese [3], NAM recognition for languages with a richer phoneme inventory such as French, using spontaneous speech and open domain is unrealistic.

In this paper, we propose two different methods to improve signal-based GMM mapping from whisper to speech. Whisper is used instead of NAM because of the difficulties in getting accurate phonetic NAM segmentation. The first method consists in using Linear Discriminant Analysis (LDA) instead of using Principal Component Analysis (PCA) to obtain the spectral vector for whisper. Classes clustering different F_0 ranges and phonemes are used for LDA. Furthermore, we compare different sizes of the context window to study the influence of spectral variation on the pitch estimation performance. In the second method, visual information is integrated as a complement to the audio information. The visual parameters are obtained by the face cloning methodology developed at ICP [11].

The paper is organized as follows. Section 2 describes some characteristics of whispered speech. Section 3 briefly describes the framework of our Whisper-to-speech conversion system already explained in [15]. Modifications in this system concerned pitch estimation. Although the results of the modifications lead to satisfactory improvements, we also investigate other direction to improve the quality of the converted speech by adding other source of information. For this reason, section 4 presents our preliminary study on the promising contribution of visual information to the conversion system proposed by Toda [14]. 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. Whispered speech, however, can only be heard by a limited set of listeners surrounding the speaker and can therefore effectively be used for quiet and private communication [7]. However, it is hard to directly use whispered speech as a medium for human communication because of its lesser intelligibility and unfamiliar perception. The conversion of whispered speech to modal voice is necessary for the realization of a "silent speech telephone".

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 the production mechanism, the acoustic characteristics of whisper differ from those of normal speech. A study on the

acoustic properties of vowels [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: 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 mainly related 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.* [4] 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 [5] indicate that "whisper pitch" is more influenced by simultaneous changes in F1 and F2 than by changes in only one of the formants.

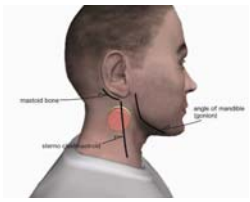

Figure 1: *Position of NAM microphone.*

2.2. NAM microphone

Nakajima *et al.* [10] proposes a new communication interface which can capture acoustic vibrations in the vocal tract from a sensor placed on the skin, below the ear (figure 1). This position 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 non-audible murmur spectral components still provide sufficient information to distinguish and recognize sound accurately [3]. Currently, the NAM microphone can record sound with frequency components up to 4 kHz while being little sensitive to external noise.

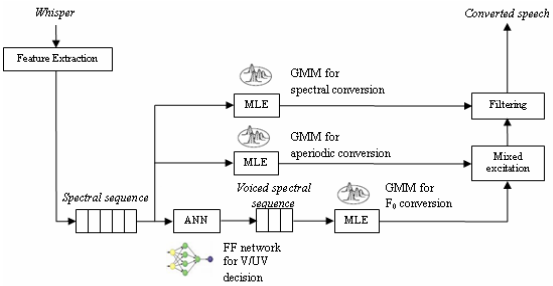

Figure 2: *Whisper-to-Speech conversion process.*

3. Using LDA for whisper-to-speech

Toda *et al.* [14] proposed a NAM-to-Speech conversion system based on GMM model [12][8] in order to convert "non-audible speech" to ordinary speech. 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. This is mainly due to impoverished synthetic intonation. In this study, we focus on improving the pitch estimation of the converted speech. We use the same schema as in the system we proposed in [15] except that the dimensionality of the spectral sequence is reduced by an LDA instead of a PCA. The diagram is shown in figure 2. In order to synthesize speech, we need to estimate not only spectral features but also excitation features, including F_0 and aperiodic components.

The spectral segment feature at each frame is constructed by concatenating spectral vectors for several frames around the current frame, in order to compensate for the impoverished phonetic features (especially for unvoiced fricatives losing their high frequency bands). Three GMMs are used to convert the segment features of whisper to three speech features, i.e., the spectrum, the F_0 and an aperiodic component which captures the noisy strength on each frequency band of excitation signal. Only voiced segments are used to train the model of F_0 estimation in order to avoid wasting Gaussian components for representing zero or undefined values of F_0 of unvoiced segments. These voiced segments are detected by a small feed-forward neural network. Estimated F_0 and aperiodic components are passed through a mixed excitation module before being combined with estimated spectra to compute the converted speech.

3.1. Evaluation

Two evaluations have been conducted, comparing this system with the system we proposed in [15].

The training corpus consists of 200 utterance pairs of whisper and speech uttered by a French male speaker and captured by a NAM microphone and a head-set microphone. The spectral characteristics of each frame are the 0^{th} through 24^{th} mel-cepstral coefficients. The context-dependent spectral feature of whispered frames are constructed by concatenating the spectral vectors at current ± 8 frames (context window). This vector is then reduced to 50 by LDA. We test the impact of the size of the context window by choosing one frame every 2, 3, 4 and 5 frames to combine with the current frame (windows varied from phoneme size ~100 ms to syllable size ~350 ms). Log-scaled F_0 characterize the target speech.

The test corpus consists of 70 utterance pairs not included in the training data which were uttered by the same speaker.

3.1.1. F_0 estimation

For this evaluation, the F_0 values of the target speech are classified into 13 classes: unvoiced frames are set to 0 Hz and voiced frames fall into 12 intervals, from 70Hz to 300 Hz. The class of a whispered frame is deduced from the class of the corresponding speech frame by aligning the two utterances. This information was then used to guide the dimension reduction of whispered vector in hopping that the relation between whispered vector and speech vector will improve the performance of the system. The number of Gaussian mixtures for F_0 estimation varies from 8 to 64 (8, 16, 32, 64). The size of the context window is also varied from the phoneme size (~100 ms) to the syllable size (~350 ms) (by picking one frame every 1-5 frames).

Table 1 shows that LDA improves the precision of pitch estimation with respect to PCA. Larger window sizes also improve the prediction. The F_0 error decreases by 16% compared to the system proposed in [15] (10.90% → 9.15%).

Figure 3 shows an example of a natural (target) F_0 curve and the synthetic F_0 curves generated by the two systems (LDA + large context window *vs.* PCA + small context window). It shows that our new system is closer to the natural F_0 curve than the old one.

Table 1. *F_0 errors (%) between converted and target speech*

method	window size (frame interval)	Number of Gaussian mixtures			
		8	16	32	64
PCA	1	10.96	10.90	10.92	**10.90**
	2	10.77	10.41	10.29	10.44
	3	10.33	9.98	10.08	10.28
	4	9.90	9.58	9.47	9.82
	5	9.44	9.17	9.32	9.31
LDA	1	10.85	10.58	10.56	10.64
	2	10.36	10.23	10.11	10.36
	3	9.98	9.94	9.93	10.29
	4	9.45	9.43	9.62	9.67
	5	**9.15**	9.22	9.25	9.37

Table 2. *Spectral distortion (dB) between converted speech and target speech*

method	window size (frame interval)	Number of Gaussian mixtures	
		8	16
PCA	1	7.23	6.96
	2	7.20	7.01
	3	7.42	7.26
	4	7.25	7.55
LDA	1	6.96	6.83
	2	6.98	7.01
	3	7.03	7.17
	4	7.19	7.34

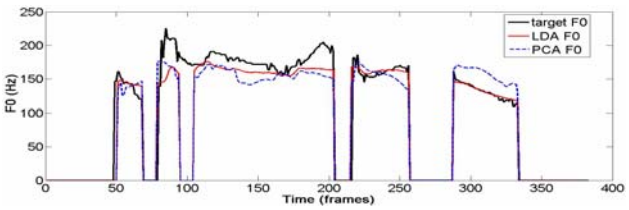

Figure 3: *Comparing natural and synthetic F_0 curves*

3.1.2. Spectral estimation

We also test the influence of LDA and long-term spectral variation to the spectral estimation. This time, we use phonetic information to get the label for whispered data to train the LDA. Each whispered frame is classified in one of 34 allophones, depending on what phoneme it belongs to.

Table 2 shows again that LDA is slightly better than PCA. Contrary to the evaluation on the F_0 estimation, when the size of the context window increases, the spectral distortion increases. In this case, the discontinuities in the input whispered vector probably degrade the performance of the system. Another plausible interpretation is that a phoneme-sized window optimally contains necessary phonetic cues for conversion.

4. Preliminary study of audiovisual whisper-to-speech conversion

To convey a message, humans produce various linguistic sounds by controlling the configuration of oral cavities. The articulators determine the resonance characteristics of the vocal tract during speech production. Therefore, speech can be characterized not only by acoustic properties but also by articulatory properties. The articulatory parameters, which vary much slower than acoustic parameters, can effectively characterize speech [13]. Some important articulators are the lips, which significantly contribute to the intelligibility of visual speech face-to-face human interaction. In the field of man and machine communication, the visual signal corresponding to speaking lips can be helpful both in input and output modalities [1]. The contribution of visual information is explored here using an accurate but unpractical lip capture system. More appropriate systems may be used in the future using wearable headset cameras.

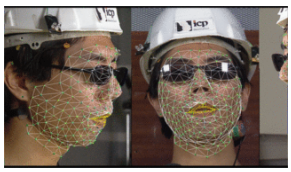

Figure 4: *Characteristic points used for capturing the movements.*

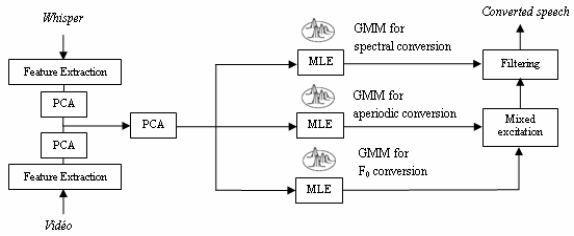

Figure 5: *Diagram of the audiovisual conversion system.*

4.1. Audiovisual conversion system

The conversion system is built using audiovisual data. The system captures, at a sampling rate of 50 Hz, the 3D positions of 142 colored beads glued on the speaker's face (Figure 4) synchrony with the acoustic signal sampled at 16000 Hz.

The shape model is built using a so-called guided Principal Component Analysis (PCA) where a priori knowledge is introduced during the linear decomposition. We compute and iteratively subtract predictors using carefully chosen data subsets [11]. For speech movements, this methodology extracts 5 components that are directly related to the rotation of the jaw, to lip rounding, upper and lower lip vertical movements and movements of the throat linked 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.

The audiovisual feature vector is obtained by combining whispered spectral and visual feature vectors in an identical way to the AAM (Active Appearance Models) introduced by Cootes [2]: each articulatory vector is multiplied with a weight *w* before concatenation with the corresponding acoustic vector. The dimension of the joint vector is further decreased by an additional PCA (see Figure 5).

4.2. Preliminary results

The database consists of 120 sentences for training and 25 sentences for the testing, pronounced by a native Japanese speaker. The 0^{th} through 19^{th} mel-cepstral coefficients are used as spectral features at each frame. The input feature vector for computing speech spectrum is constructed by concatenating feature vectors at current ±8 frames and further reduced to a 40-dimention vector by a PCA. Similarly to the processing of the acoustic signal, each visual frame is interpolated at 200 Hz – so as to be synchronous with the audio processing – and characterized by a feature vector obtained by concatenating and projecting ±8 frames centered around the current frame on the first n principal components. The dimension of the visual vector n is set to 10, 20, 40 or 50. The

Table 3. *Contribution of visual information to (a) spectral estimation; (b) voiced/unvoiced decision; (c) F0 estimation. Best performance (bold) is obtained for a balanced contribution of audio and visual parameters.*

		Visual weight									
Distorsions	Visual dimension	**Audio-only**	0.25	0.5	0.75	**1**	1.25	1.5	1.75	2	**Video-only**
(a) Cepstral distortion in dB	10		5.66	5.63	5.61	5.58	5.60	5.63	5.62	5.65	
	20		5.68	5.63	5.60	**5.56**	5.60	5.61	5.60	5.62	
	40	**5.69**	5.68	5.63	5.60	5.61	5.57	5.61	5.60	5.62	**9.89**
	50		5.68	5.63	5.60	5.60	5.57	5.61	5.59	5.62	
(b) Voiced/ unvoiced detection (%)	10		13.79	13.48	12.90	12.67	20.71	20.67	20.97	21.22	
	20		13.24	13.58	12.56	**12.36**	20.73	20.28	20.57	19.53	
	40	**14.81**	13.24	13.58	12.56	12.70	20.45	20.53	20.36	20.26	**31.34**
	50		13.24	13.58	12.56	13.38	20.45	20.51	20.74	20.26	
(c) F0 estimation (%).	10		18.39	18.14	17.54	17.27	24.58	24.52	24.77	25.53	
	20		17.85	18.21	**17.14**	17.47	24.62	24.15	24.51	23.53	
	40	**19.48**	17.85	18.21	17.14	17.28	24.39	24.41	24.47	24.21	**36.31**
	50		17.85	18.21	17.14	17.93	24.39	24.37	24.63	24.21	

weight *w* was also changed from 0.25 to 2. The conversion system uses the first 40 principal components of joint audiovisual vector. In this evaluation, the number of Gaussian was fixed at 16 for the spectral estimation, 8 for the F_0 estimation and 8 for the aperiodic components estimation.

Table 3 shows the positive contribution of visual information on the performance of the conversion. The best results are obtained with $w = 1$ and a dimension of the visual vector of 20. The spectral distortion between the converted speech and the modal speech is decreased by 2.3% while the error decreases by 16.5 % for voiced/unvoiced detection and 10.3% for F_0 estimation. With visual information only, the performance of the system is significantly degraded.

5. Conclusions

This paper describes our modifications to improve the intelligibility and the naturalness of the converted speech of the whisper-to-speech system based on GMM model. First, the use of LDA with a large context window significantly improved the converted speech, compared with using PCA with a small window. Secondly, the preliminary results on the contribution of visual information on a Japanese corpus encourage us to continue in this direction using a larger audiovisual corpus. Although the performance of the system is improved and the difference is clearly audible, the estimated pitch is still too flat due to the GMMs. In the future, we will investigate how to obtain audible speech from whisper by using a HMM which is more appropriate for modelling a time sequence of speech parameters.

Acknowledgements

The authors are grateful to C. Vilain, C. Savariaux, A. Arnal, K. Nakamura & T. Toda for data acquisition, to T.Toda for letting us use the NAM system, to Prof. Hideki Kawahara of Wakayama University in Japan for the permission to use the STRAIGHT system.

References

[1] Campbell, R., B. Dodd, and D. Burnham, *Hearing by Eye II. The Psychology of Speechreading and Auditory-Visual Speech*. 1998, Hove, UK: Psychology Press Ltd.

[2] Cootes, T.F., Edwards, G.J. and Taylor C.J. "Active Appearance Models", IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6):681-685, 2001.

[3] Heracleous, P. et al., "A tissue-conductive acoustic sensor applied in speech recognition for privacy", International Conference on Smart Objects & Ambient Intelligence, Grenoble-France, 93-98, 2005.

[4] Higashikawa, M., Nakai, K., Sakakura, A. and Takahashi, H., "Perceived Pitch of Whispered Vowels – Relationship with formant frequencies: A preliminary study", 155-158.

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

[6] Hueber, T. et al., "Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips", Interspeech, 2007.

[7] Ito, T., Takeda, K. and Itakura, F., "Analysis and recognition of whispered speech", Speech Communication, Lisboa, 45(2), 139-152, 2005.

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

[9] Nakagiri, M. et al., "Improving body transmitted unvoiced speech with statistical voice conversion", Interspeech, Pittsburgh, 2270-2273, 2006.

[10] Nakajima, Y., et al., "Non-audible murmur recognition", Eurospeech, Geneva, Switzerland, 2601-2604, 2003.

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

[12] 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.

[13] Toda, T., Black, A.W. and Tokuda, K., "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model", Speech Communication, 50, 215-227, 2008.

[14] Toda, T. and Shikano, K., "NAM-to-Speech conversion with gaussian mixture models", Interspeech, Lisbon-Portugal, 1957-1960, 2005.

[15] Tran, V-A., Bailly, G., Loevenbruck, H. and Toda, T., "Predicting F_0 and voicing from NAM-captured whispered speech", Speech Prosody, 2008.