Improvement to a NAM-captured whisper-to-speech system

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

1. GIPSA-lab, UMR 5216 CNRS/Grenoble Universities, France
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

Exploiting a tissue-conductive sensor – a stethoscopic microphone – the system developed at NAIST which converts Non-Audible Murmur (NAM) to audible speech by GMM-based statistical mapping is a very promising technique. The quality of the converted speech is however still insufficient for computer-mediated communication, notably because of the poor estimation of F_0 from unvoiced speech and because of impoverished phonetic contrasts. This paper presents our investigations to improve the intelligibility and naturalness of the synthesized speech and first objective and subjective evaluations of the resulting system. The first improvement concerns voicing and F_0 estimation. Instead of using a single GMM for both, we estimate a continuous F_0 using a GMM, trained on target voiced segments only. The continuous F_0 estimation is filtered by a voicing decision computed by a neural network. The objective and subjective improvement is significant. The second improvement concerns the input time window and its dimensionality reduction: we show that the precision of F_0 estimation is also significantly improved by extending the input time window from 90 to 450ms and by using a Linear Discriminant Analysis (LDA) instead of the original Principal Component Analysis (PCA). Estimation of spectral envelope is also slightly improved with LDA but is degraded with larger time windows. A third improvement consists in adding visual parameters both as input and output parameters. The positive contribution

of this information is confirmed by a subjective test. Finally, HMM-based conversion is compared with GMM-based conversion.

Keywords: non-audible murmur, whispered speech, audiovisual voice conversion, silent speech interface.

1. Introduction

In recent years, advances in wireless communication technology have led to the widespread use of cellular phones for speech communication. Because of noisy environmental conditions and competing surrounding conversations, users tend to speak loudly. As a consequence, private policies and public legislation tend to restrain the use of cellular phones in public places. Silent speech which can only be heard by a limited set of listeners close to the speaker is an attractive solution to this problem if it can efficiently be used for quiet and private communication. Silent speech capture interfaces have already been developed including electromyography (Bett and Jorgensen 2005; Jorgensen and Binsted 2005; Jou, Schultz *et al.* 2006; Walliczek, Kraft *et al.* 2006), non-audible murmur microphone (Heracleous, Nakajima *et al.* 2005; Toda and Shikano 2005), ultrasound and optical imagery (Hueber, Aversano *et al.* 2007; Hueber, Chollet *et al.* 2007; Hueber, Chollet *et al.* 2007; Hueber, Chollet *et al.* 2008; Hueber, Chollet *et al.* 2008) together with several techniques to convert silent speech signals to audible speech.

Two possible ways to map silent speech to modal speech have been proposed: (a) Combining speech recognition and synthesis techniques as proposed by Hueber *et al.* in the Ouisper project (Hueber, Chollet *et al.* 2007). By introducing linguistic levels both in the recognition and synthesis, such systems can potentially compensate for the impoverished input by including linguistic knowledge into the recognition process. The quality of the output speech is either excellent or extremely degraded depending on recognition performance. (b) Direct signal-to-signal mapping using aligned corpora (Toda and Shikano 2005). By using the NAM microphone to capture non-audible murmur, Toda *et al.* proposed a NAM-to-speech conversion system based on a GMM

model in order to convert "non-audible" speech to audible speech. It was shown that this system provides intelligible speech with constant quality but the naturalness of the converted speech is still unsatisfactory. This is due to the poor F_0 estimation from unvoiced speech. Note that the F_0 estimation problem is a difficulty shared by all the systems described above, either using whispered speech, or non-audible murmur. Although whispered speech typically does not involve any vocal fold vibration, laryngeal activity may however exist during whispered speech, and this could represent useful information for the F_0 estimation process. Coleman *et al.* (2002) have shown, using dynamic Magnetic Resonance Imaging, that larynx movements related to intonation changes (rising and falling pitch) can be observed in whispered speech. According to the authors this offers an explanation for perceived pitch in whispered speech: laryngeal movements modify the shape of the oral cavity, thus altering the vocal tract acoustics so that pitch changes can be inferred. Subvocal speech, such as the murmur that can sometimes be observed in hallucinating schizophrenic patients, has also been shown to be associated with laryngeal activity. Inouye and Shimizu (1970) reported increased EMG activity in speech-related muscles including laryngeal muscles (cricothyroid, sternohyoid, orbicularis oris, and depressor anguli oris) in 47.6% of the hallucinations of nine schizophrenic patients. These works suggest that whispered speech might carry information on laryngeal activity. This activity could be useful in the recovery of F_0 because it could modify the shape of the oral cavity and hence be audible. It has been shown by Higashikawa *et al.* (1996) that when speakers try to whisper high pitch vowels vs low pitch vowels, F1 and F2 are raised and listeners are able to identify the pitch correctly. This formant frequency raising could be due to laryngeal movement. Other sources of pitch information may exist. Zeroual *et al.* (2005) and Crevier-Buchman *et al.* (2008) have shown with ultra- high-speed cinematography that whisper is associated with an anterior-posterior epilaryngeal compression and an abduction of the ventricular bands. These supraglottic changes may also be sources of information on F_0 during whisper.

In this article, we propose two major improvements to the original signal-based GMM mapping from whisper to speech proposed by Toda *et al.* As advocated by Nakagiri *et al.* (2006), whisper is used here instead of NAM because of the difficulty of getting accurate phonetic NAM segmentation. The first improvement concerns the input feature space. The second improvement consists in incorporating visual information into both the input and output feature spaces. The visual parameters are obtained by the face cloning methodology developed at Gipsa-Lab (Revéret, Bailly *et al.* 2000; Badin, Bailly *et al.* 2002; Bailly, Bérar *et al.* 2003).

Section 2 gives a brief description of the NAM microphone. Section 3 describes the original NAM-to-speech system proposed by Toda *et al.* Section 4 then describes our improvement for the performance of this system, especially for the naturalness of the converted speech while section 5 focuses on the contribution of visual features. Session 6 is dedicated to the testing of a HMM-based conversion system. Finally, a discussion of our results is provided in section 7.

2. NAM microphone

The tissue-conductive microphone proposed by Nakajima *et al.* (2003) comprises an electret condenser microphone (ECM) covered with a soft polymer material, such as soft silicone or urethane elastomer, which provides better impedance matching with the soft tissue of the neck. This microphone can capture acoustic vibrations in the vocal tract from a sensor placed on the skin, below the ear (Figure 1). This position allows a high quality recording of various types of body-transmitted speech utterances, such as normal speech, whisper and non-audible murmur, even in the environmental noises. 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. The soft silicone NAM microphone used in our laboratory can record sound with frequency components up to 3 kHz while being little sensitive to external noise. Figure 2 shows an example of a whispered utterance in French, captured by this microphone.

3. Original NAM-to-Speech conversion system

To characterize whisper, namely unvoiced speech without vocal fold vibration, spectral envelope and power information are used, while spectral envelope, power, F_0 and aperiodic components are used for speech synthesis using STRAIGHT, a versatile speech vocoder developed by Kawahara *et al.* (1999) .

Before training the models for spectrum and excitation estimation, pairs of whisper and speech uttered by a speaker must be aligned, in order to determine corresponding frames. Phonetic transcription information were used to get a better alignment compared to automatic Dynamic Time Warping (DTW).

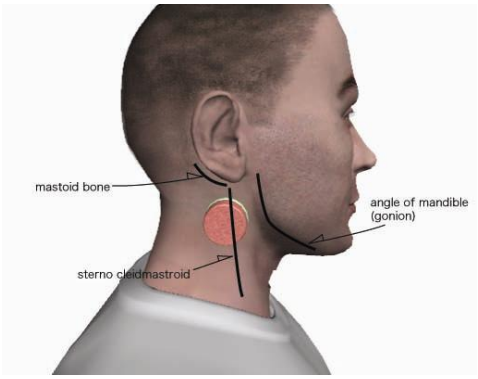


Figure 1 : Position of the NAM microphone

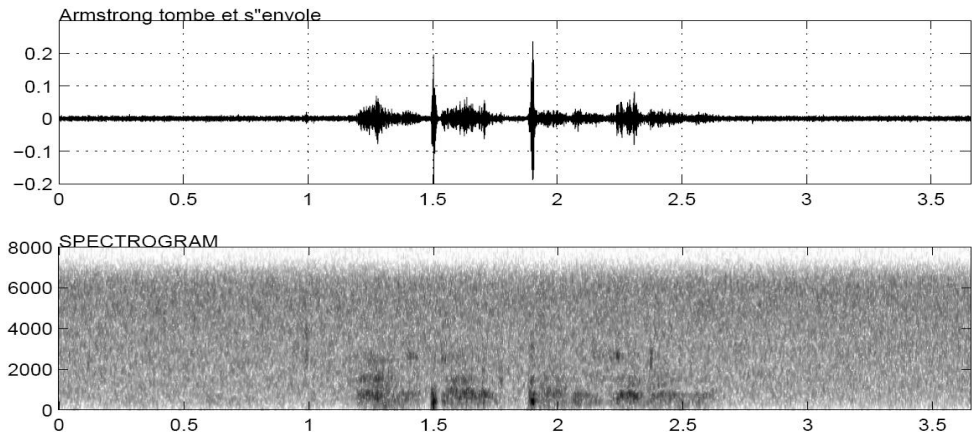


Figure 2 : Whispered speech captured by a NAM microphone for the French utterance: "Amstrong tombe et s'envole".

3.1. Spectral estimation

We use the same schema for spectral estimation as the one proposed by Toda *et al.* (Toda and Shikano 2005). As described in Toda *et al.* (2009), let us use an input static feature $x_t = x_t(1),...,x_t(D_x)^{\mathrm{T}}$ and an output static feature vector $y_t = y_t(1),...,y_t(D_x)^{\mathrm{T}}$ at frame *t*, respectively. In order to compensate for the lost characteristics of some phonemes due to body transmission, we use a segment feature $X_t = W_x\left[x_{t-L}^{\mathrm{T}},...x_t^{\mathrm{T}},x_{t+L}^{\mathrm{T}}\right]^{\mathrm{T}} + b_x$ extracted over several frames ($t \pm L$) as an input speech parameter vector where W_x and b_x are determined by PCA in this paper. As an output speech parameter vector, we use $Y_t = \left[y_t^{\mathrm{T}}, \Delta y_t^{\mathrm{T}}\right]$ consisting of both static and dynamic feature vectors.

Using parallel training data set consisting of time-aligned input and output parameter vectors $\left[X_1^{\mathrm{T}},Y_1^{\mathrm{T}}\right]^{\mathrm{T}}$, $\left[X_2^{\mathrm{T}},Y_2^{\mathrm{T}}\right]^{\mathrm{T}}$, …, $\left[X_T^{\mathrm{T}},Y_T^{\mathrm{T}}\right]^{\mathrm{T}}$, the joint probability density of the input and output parameter vectors is modelled by a GMM as follows:

$$P(X_t,Y_t|\lambda) = \sum_{m=1}^{M} w_m \mathrm{N}([X_t^{\mathrm{T}},Y_t^{\mathrm{T}}]^{\mathrm{T}};\mu_m^{(X,Y)},\Sigma_m^{(X,Y)})$$

$$\mu_m^{(X,Y)} = \begin{bmatrix}\mu_m^{(X)} \\ \mu_m^{(Y)}\end{bmatrix},\ \Sigma_m^{(X,Y)} = \begin{bmatrix}\Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)}\end{bmatrix}$$

where $N(\mu, \sum)$ denotes Gaussian distribution with a mean vector μ and a covariance matrix $\sum$. The mixture component index is *m*. The total number of mixture components is *M*. A parameter set of the GMM is λ, which consists of weights w_m, mean vectors $\mu_m^{(X,Y)}$ and full covariance matrices $\Sigma_m^{(X,Y)}$ for individual mixture components.

The probability density of the GV of the output static feature vectors over an utterance is also modelled by a Gaussian distribution,

$$P(v(y)|\lambda^{(v)}) = N(v(y); \mu^{(v)}, \Sigma^{(v)})$$

where the GV $v(y) = \left[v(1), ..., v(D_y) \right]^{\mathrm{T}}$ is calculated by

$$v(d) = \frac{1}{T} \sum_{t=1}^{T} \left(y_t(d) - \frac{1}{T} \sum_{\tau=1}^{T} y_\tau(d) \right)^2$$

A parameter set $\lambda^{(v)}$ consists of a mean vector $\mu^{(v)}$ and a diagonal covariance matrix $\Sigma^{(v)}$. This global variance information is used to reduce the over-smoothing, which is inevitable in the conventional ML-based parameter estimation (Toda and Tokuda 2005).

In the conversion procedure, let $X = \left[X_1^{\mathrm{T}}, ..., X_t^{\mathrm{T}}, ..., X_T^{\mathrm{T}} \right]^{\mathrm{T}}$ and $Y = \left[Y_1^{\mathrm{T}}, ..., Y_t^{\mathrm{T}}, ..., Y_T^{\mathrm{T}} \right]^{\mathrm{T}}$ be the time sequence of the input parameter vectors and that of the output parameter vectors, respectively. The converted static feature sequence $y = \left[y_1^{\mathrm{T}}, ..., y_2^{\mathrm{T}}, ..., y_T^{\mathrm{T}} \right]^{\mathrm{T}}$ is determined by maximizing a product of the conditional probability density of *Y* given *X* and the GV probability density as follows:

$\hat{y} = \arg\max_{y} P(Y|X, \lambda)^w P(v(y)|\lambda^{(v)})$ subject to $Y = W_y y$ where W_y is a window matrix to extend the static feature vector sequence to the parameter vector sequence consisting of static and dynamic features.

3.2. Excitation estimation

The speech excitation is decomposed into two components: a periodic or quasi-periodic component which takes into account the quasi-periodic segments of speech produced by the regular vibrations of the vocal folds and an aperiodic component for the noise (frication, aspiration, bursts). In the STRAIGHT system (Kawahara, Masuda-Katsuse *et al.* 1999), the mixed excitation is defined as the frequency-dependent weighted sum of these two components. The weight is determined based on an aperiodic component in each frequency band which is calculated from the smoothed power spectrum by a subtraction of the lower spectral envelope from the upper spectral envelope. The

upper envelope is calculated by connecting spectral peaks and the lower envelope is calculated by connecting spectral valleys.

Two GMM models for F_0 estimation and aperiodic estimation are constructed in the same way as the spectral estimation except that the global variance (GV) is not used because GV does not cause large differences to the converted speech in the aperiodic conversion. Static and dynamic features Y_t of F_0 and aperiodic components are used while keeping the same feature vector of whisper X_t as that used for the spectral estimation.

4. Improvement of intelligibility and naturalness of the converted speech

This section is described in more details in Tran *et al.* (2008).

4.1. Feature extraction

The training corpus consists in 200 utterance pairs of whisper and speech uttered by a native male speaker of French and captured by a NAM microphone and a 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 are used as spectral features at each frame. The spectral segment features of whisper are constructed by concatenating feature vectors at each current whispered frame ± 8 frames (i.e. representing 90 ms of speech) to take into account the context. Then the vector dimension is reduced to 50 using a PCA technique. Log-scaled F_0 extracted with fixed-point analysis and 5 average dB values of the aperiodic components on five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6, 6 to 8 kHz) are used to characterize the excitation feature.

The test corpus consists of 70 utterance pairs not included in the training data.

4.2. Voicing decision

In the original system, F_0 and voicing are jointly estimated by a unique GMM model: unvoiced frames are assigned with $F_0 = 0$ for training and voicing decision is then determined using a simple

threshold. The main drawback of this approach is to waste Gaussian components to represent zero values of F_0 for unvoiced segments and its danger is to predict unstable F_0 values for unvoiced segments incorrectly labelled as voiced. In order to improve this joint representation, we only use voiced segments to train the GMM model for F_0. For synthesis, a feed-forward neural network is used to predict the segments from the whispered speech. This network has 50 input cells (for the 50 dimension feature vector), 17 hidden cells, and 1 continuous output. The continuous output is converted to a binary output by applying a threshold. When the output is larger than .5, the frame is labelled as voiced. Otherwise, it is labelled as unvoiced. Only voiced frames are used in the GMM. Continuous F_0 values are then predicted, from the voiced values. Figure 3 presents the synopsis of the system.

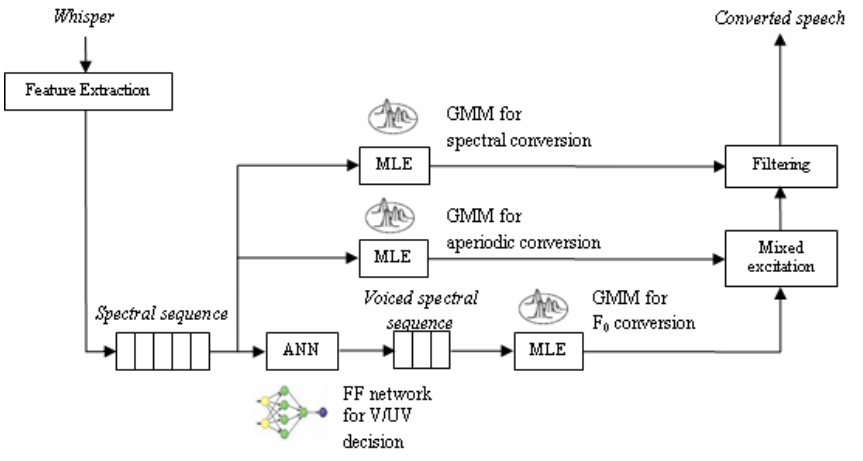


Figure 3. Synopsis of the whisper-to-speech conversion method

Table 1 shows the detection performance of this network in comparison with the error in the original system. We have a significant improvement of the voiced/unvoiced detection.

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

4.3. F_0 estimation

As stated above, only voiced segments in whispered speech were used to train the GMM model for the F_0 estimation. We also compared our system with the original one with different number of mixtures on both the training and the test data. Full covariance matrices were used for both GMMs. The difference 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. This difference is given by the following formula: $Diff = (synthetic\_F_0 - natural\_F_0)/ natural\_F_0$. Figure 4 shows that the proposed framework outperforms the original system. In addition, when the number of Gaussian mixtures increases, the errors of both systems on the training data decrease, but the errors on test data are not sensitive to the number of mixtures.

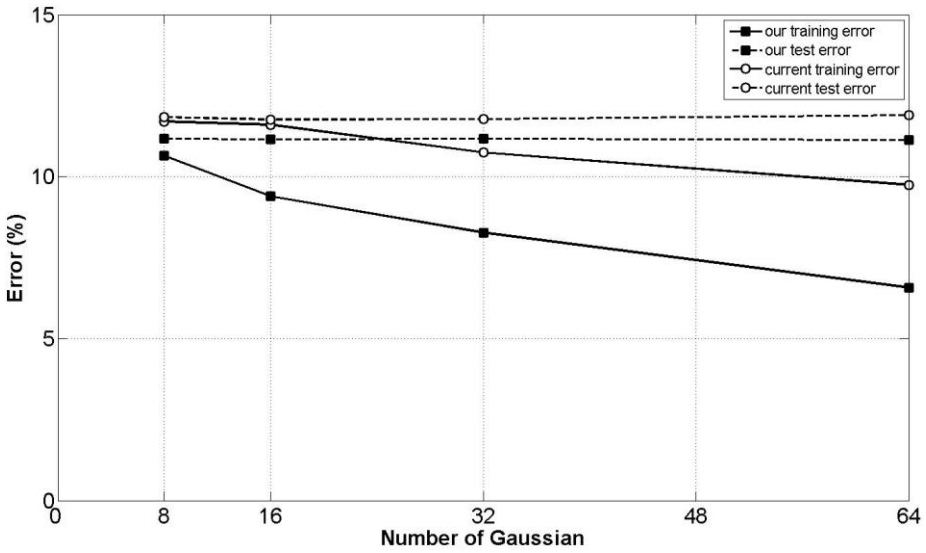


Figure 4. Parameter evaluation on training (solid line) and test corpus (dashed line).

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 proposed system is closer to the natural F_0 curve than the original system, especially for the final contour.

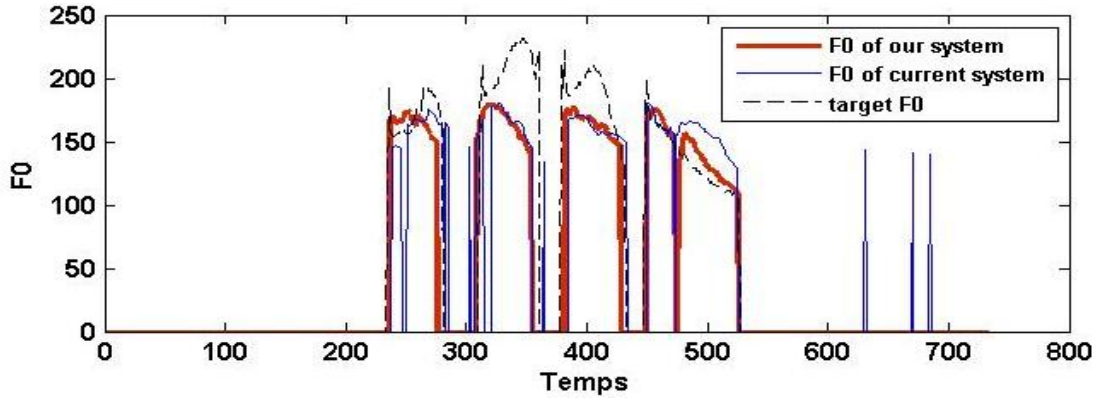


Figure 5. Natural and synthetic F_0 curve for the utterance: "Armstrong tombe et s'envole".

4.4. Perceptual Evaluation

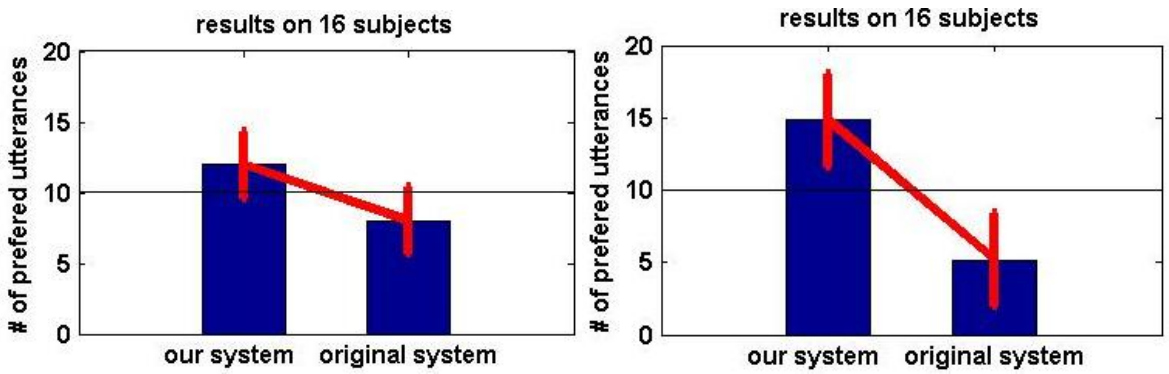
Sixteen French listeners who had never listened to NAM participated in our perceptual tests on intelligibility and naturalness of the converted speech from the two systems. We used 20 test utterances not included in the training set.

Each listener heard an utterance pronounced in modal speech and the converted utterances obtained from the whispered speech with both systems. For intelligibility testing, all parameters were obtained by voice conversion and synthesis was performed using STRAIGHT (Kawahara, Masuda-Katsuse *et al.* 1999). For naturalness, the spectral parameters were original and the stimuli were obtained by substituting only predicted voicing and F_0 values to the original warped target frames.

This procedure was chosen because the quality of the converted spectrum could also influence the perception in the naturalness test. We wanted to make sure that we were testing F0 and voicing alone (not other potentially influential parameters).

For each utterance, listeners were asked which one was closer to the original one, in terms of intelligibility and in terms of naturalness. An ABX test (or matching-to-sample test) was used. It is a discrimination procedure involving presentation of two test items and a target item. The listeners are asked to tell which test item (A or B) is closest to the target (X). Figure 6.a displays the mean intelligibility scores for all the listeners for the converted sentences using the original system and

our new system. An ANOVA showed that the intelligibility score is higher for the sentences obtained with our new system ($F = 23.41$, $p < .001$). Figure 6.b shows the mean naturalness. Again the proposed system is strongly preferred to the original one, as shown by an ANOVA ($F = 74.89$, $p < .001$).



(a) (b)

Figure 6. ABX results. (a) intelligibility ratings (b) naturalness ratings

4.5. Influence of time window and LDA

4.5.1. F0 estimation

For this evaluation, the whispered frames are classified into 13 classes: unvoiced frames are labelled with '0' label and voiced frames fall into 12 other labels, depending on which interval the F_0 value in this frame belongs to (bark scale between 70Hz and 300 Hz). The class of a whispered frame is deduced from the class of the corresponding speech frame using the warp path. The number of Gaussian mixtures for F_0 estimation varied from 8 to 64 (8, 16, 32, 64). The size of the context window was also varied from the phoneme size (90 ms) to the syllable size (450 ms) (by picking one frame every 1-5 frames).

Table 2 shows that using LDA and a large window size improves the precision of pitch estimation with respect to PCA with a small window. When using LDA with a window of 5, the F_0 error

decreases by 16% compared to our previous system with PCA and a small window size (10.90% → 9.15%). However, using LDA instead of PCA with the same context window size does not significantly improve the precision (9.17% → 9.15%).

Table 2. F_0 difference (%) 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

Figure 7 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). The predicted F_0 contour is closer to the natural F_0 curve than the one generated by the reference system and also smoother.

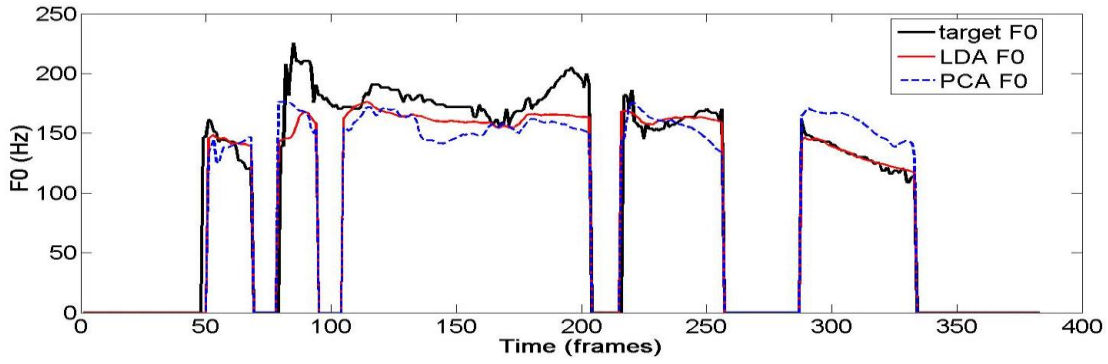


Figure 7. Comparing natural and synthetic F_0 curves

4.5.2. Spectral estimation

We also evaluated the influence of LDA and long-term spectral variation to the spectral estimation. The phonetic segmentation was used here to perform the LDA: each whispered frame was classified into one of 34 classes, depending on which phonetic segment it belonged to.

Table 3 provides the cepstral distortion between the converted and the target speech (the higher the distortion, the worse the performance). It shows that LDA is slightly better than PCA. But contrary to F_0 estimation, the spectral distortion increases when the size of the time window increases. The most plausible interpretation is that a phoneme-sized window optimally contains necessary contextual cues for spectral conversion.

Table 3. Cepstral distortion (dB) between converted 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

5. Audiovisual conversion

To convey a message, humans produce sounds by controlling the configuration of oral cavities. The speech articulators determine the resonance characteristics of the vocal tract during speech production. Movements of visible articulators such jaw and lips are known to significantly contribute to the intelligibility of speech during face-to-face communication (Summerfield 1979; Summerfield, MacLeod *et al.* 1989). In the field of person-machine communication, the visual

information can be helpful both as input and output modalities (Bailly, Bérar *et al.* 2003; Potamianos, Neti *et al.* 2003).

5.1. Audiovisual corpus

An audiovisual conversion system was built using audiovisual data pronounced by a native Japanese speaker. 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 8) in synchrony with the acoustic signal sampled at 16000 Hz.

Figure 8. Characteristic points used for capturing the movements. The mesh used to capture and train the texture model is superimposed to the original video-captured images.

5.2. Visual parameters extraction and shape modelling

A 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 (Badin, Bailly *et al.* 2002). For speech movements and for each particular speaker and language, 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 that are 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.

5.3. Contribution of visual parameters for the conversion system

For this evaluation, the database consists of 150 sentences for training and 40 sentences for testing, pronounced by a native Japanese speaker. The audiovisual feature vector combines whispered spectral and visual feature vectors in an identical way to the AAM (Active Appearance Models) introduced by Cootes (Cootes, Edwards *et al.* 2001) where two distinct dimensionality reductions by PCA are performed in shape and appearance and further combined into a third PCA to get a joint shape and appearance model. Acoustic and visual features are fused here, using an identical process. 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 centred around the current frame on the first *n* principal components. The dimension of the visual vector *n* is set to 10, 20 or 40. Prior to joint PCA, the visual vector is weighted. The weight *w* was also changed from 0.25 to 2. The conversion system uses the first 40 principal components of this joint audiovisual vector. In the evaluation, the number of Gaussian is fixed at 16 for the spectral estimation, 16 for the F_0 estimation and 16 for the aperiodic components estimation.

Table 4. Influence of visual information on voicing decision.

Type of error	Feed-fwd NN (%)					GMM (%)
	AU	AUVI				
		w = 1	w = 0.75	w = 0.5	w = 0.25	
Voiced error	3.71	3.58	3.74	3.34	**2.75**	4.29
Unvoiced error	4.34	**4.19**	4.71	4.37	4.87	5.47
Total	8.05	7.77	8.45	7.71	**7.62**	9.76

5.3.1. Voicing decision

In the same way as for the evaluation of voicing decision described in section 4.2, the audiovisual vectors of whisper in the training corpus of the conversion system are used to train the network. This network has 40 input neurons, 17 hidden neurons and 1 output neuron. Table 4 shows that visual information improves the accurateness of voicing decision. With a visual weight empirally set at 0.25, the voicing error is decreased by 5.4 % (8.05 % $\rightarrow$ 7.62%) compared to audio-only and 21.9 % compared with the baseline system (9.76 % $\rightarrow$ 7.62%).

Table 5. Influence of visual information on the estimation of spectral and excitation features.

		Visual weight									
Distorsions	Visual dimension	**Audio-only**	0.25	0.5	0.75	**1**	1.25	1.5	1.75	2	**Video-only**
Cepstral distortion (dB)	10	**5.99**	7.01	5.97	5.80	5.78	5.88	5.87	5.86	5.84	**9.28**
	20		7.01	5.97	5.79	5.77	5.86	5.83	5.95	5.89	
	40		7.02	5.96	5.79	**5.77**	5.86	5.82	5.94	5.88	
F_0 estimation (%).	10	**11.56**	14.55	12.44	11.42	10.99	11.71	12.75	14.24	15.77	**12.95**
	20		14.79	12.38	11.40	10.78	11.66	12.62	14.31	15.43	
	40		14.57	12.39	11.04	**10.76**	11.66	12.62	14.33	15.46	
AP distortion (dB)	10	**38.72**									**51.45**
	20		41.42	38.25	37.79	37.55	37.33	38	37.90	37.95	
	40		41.42	38.21	37.78	37.53	**37.33**	37.99	37.53	37.93	

5.3.2. Spectral and excitation estimations

The visual information also enhances the performance of the conversion system. As shown in Table 5, the best results are obtained with weighting equally acoustic and visual parameters (w = 1) and with the dimension of the visual vector equal to 40. The spectral distortion between the converted speech and the target speech is decreased by 3.7% (5.99 dB $\rightarrow$ 5.77 dB) while the difference between the converted speech and that of target speech is decreased by 6.9 % (11.56 % $\rightarrow$ 10.76 %) for F_0 estimation and 3.6 % for aperiodic components estimation.

Figure 9 shows an example of F_0 curves converted from audio and audiovisual input whisper. With visual information as an additional input, we have a better converted F_0.

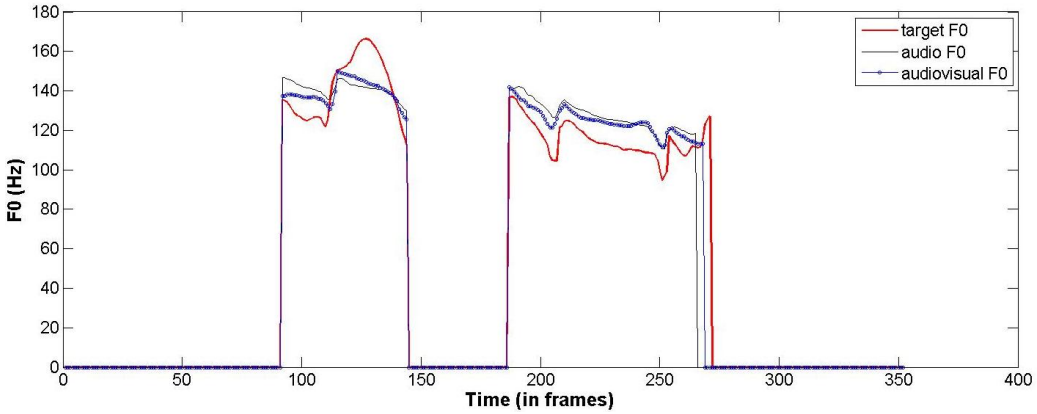


Figure 9. Natural and synthetic F_0 curve converted from audio and audiovisual whisper

5.3.3. Facial animation

The audiovisual rendering of estimated visual parameters is performed by texturing the mesh piloted by the shape model introduced in section 5.2. An appearance model that computes a texture at each time frame is built using a technique similar to AAM (Cootes, Edwards *et al.* 2001) in three steps: (1) shape-free (SF) images are obtained by morphing key images to a reference mesh; (2) contrary to AAM, these pooled SF images are then directly linked to articulatory parameters via a multilinear regression; (3) the resulting appearance model is then used to compute a SF synthetic image given the set of articulatory parameter of each frame and used to texture the corresponding shape. The main difference with AAM is the direct multilinear regression used instead of a joint PCA and the number of configurations used: while AAM typically uses a few dozen images on which a generic mesh is adapted by hand or semi-automatically, we use here more than a thousand configurations on which the mesh is positioned automatically thanks to marked fleshpoints (Bailly, Bégault *et al.* 2008). The videorealistic audiovisual rendering of computed facial movements is illustrated in Figure 10.

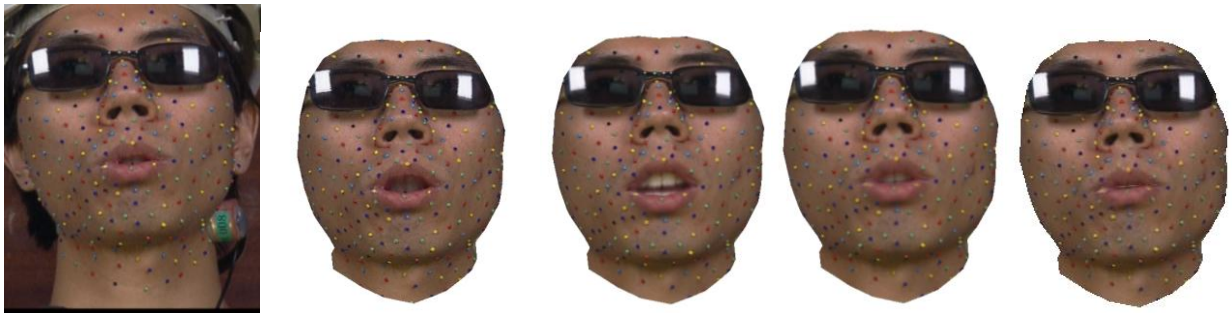


Figure 10. Video-realistic rendering of computed movements by statistical shape and appearance models driven by the same articulatory parameters. From left to right: original frame; its shape-free transform (note the texture distortion in the mouth region because the reference mesh is chosen with an open mouth) and three different synthesized frames sampling a bilabial closure (note the nice rendering of the inner mouth despite the linear modelling of the nonlinear appearance/disappearance of the inner parts, notably the teeth).

5.3.4. Subjective evaluation

Eight Japanese listeners participated in our perceptual tests on audiovisual converted speech. The stimuli consisted of Japanese VCV (with V chosen amongst five vowels and C amongst twenty-seven consonants) sequences with four conditions:

1. Speech generated from whispered audio (named ‘condition A from A’)
2. Speech generated from whispered audio and video (‘condition A from AV’)
3. Speech and facial animation generated from whispered audio (‘condition AV from A’)
4. Speech and facial animation generated from whispered audio and video (‘condition AV from AV’)

The five vowels were /a/, /i/, /e/, /o/, /u/. The 27 consonants were the following: /p/, /pj/, /b/, /bj/, /m/, /mj/, /d/, /t/, /s/, /ts/, /z/, /j/, /n/, /nj/, /k/, /kj/, /g/, /gj/, /f/, /ʃ/, /tʃ/, /ʒ/, /h/, /hj/, /r/, /rj/, /w/. Only 115 combinations of vowels and consonants were tested.

Each participant heard and viewed a list of randomized synthetic audio and audiovisual VCV. For each VCV, she/he was asked to select what consonant she/he heard among a list of 27 possible Japanese consonants. Figure 11 provides the mean recognition scores for all the participants. Visual parameters significantly improve consonant recognition: the identification ratios for ‘AV from A’ (28.37 %), ‘A from AV’ (30.56 %) and ‘AV from AV’ (36.21 %) are all significantly higher than that of the ‘A from A’ condition (23.84 %) ($F = 1.23$, $p < 0.303$). The figure also shows that

providing visual information to the speakers is more beneficial when it is synthesized from audiovisual data ('AV from AV') than when it is derived from audio data alone ('AV from A'). Furthermore the addition of visual information in the input data ('A from AV') increases identification scores compared with audio data alone ('A from A'), but is not significantly different from providing audiovisual information derived from the audio alone ('AV from A').

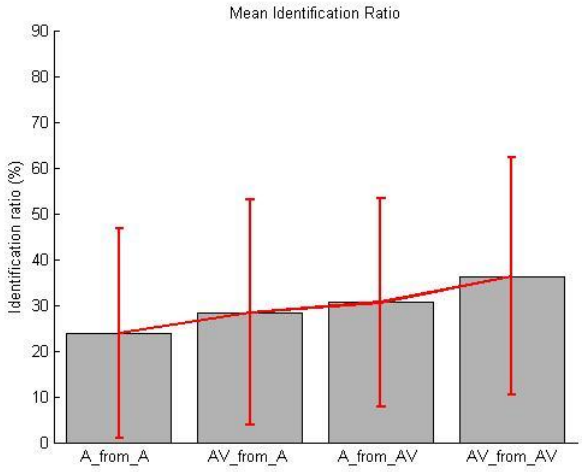


Figure 11. Mean Recognition ratio (F = 1.23, p < 0.303).

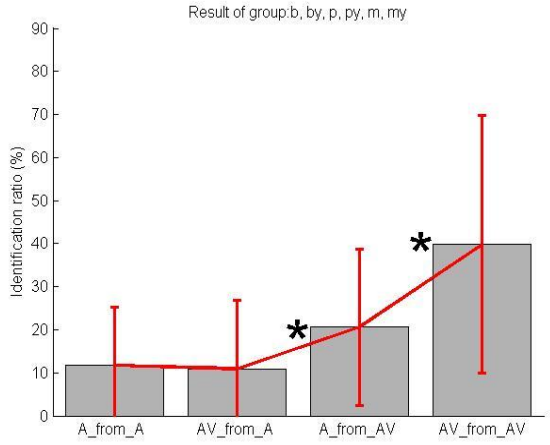
Although adding visual information greatly increases the identification scores on all the tested VCVs, the scores are still quite low (less than 40%). It should be noted however that nonsense VCV recognition is a difficult task, especially in a language with a lot of consonants. It could therefore be argued that identification scores on words or sentences could be much higher, with the help of lexical, syntactic and contextual information. With this in mind, it could be interesting to check whether the addition of visual information provided in fact some cues to place of articulation, even if accurate phoneme detection was too difficult. Therefore we also grouped the consonants into different place of articulation categories to examine which consonantal groups benefited most from the addition of visual information. The 27 consonants were thus grouped into bilabials (/p/, /pj/, /b/, /bj/, /m/, /mj/), alveolars (/d/, /t/, /s/, /ts/, /z/, /j/, /n/, /nj/), palatals (/k/, /kj/, /g/, /gj/), non-alveolar fricatives (/f/, /ʃ/, /tʃ/, /ʒ/) and others (/h/, /hj/, /r/, /rj/, /w/). The first aim was to check whether

consonants belonging to the bilabial category, which are intrinsically more salient visually, were better identified as bilabial (be it /p/, /pj/, /b/, /bj/, /m/ or /mj/), when visual information was taken into account in the speech conversion procedure and when audiovisual stimuli were presented. The second aim was to examine what kind of perceptual confusions were observed.

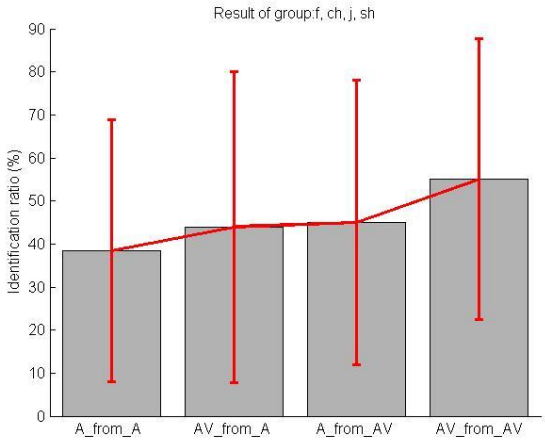
Figure 12 displays the evolution of the identification scores along the four conditions ('A from A', 'AV from A', 'A from AV', 'AV from AV') for the five articulatory groups. This articulatory grouping shows that visual information significantly helps the participants to identify bilabials. The identification ratio for the bilabial consonants rises from 11.67 % in the 'A from A' condition to 39.83 % in the 'AV from AV' condition ($F = 2.62$, $p < 0.079$). The identification ratios for all the other groups are quite stable and are not significantly increased by the addition of visual information.

Table 6 provides the confusion matrices with places of articulation chosen among 5 categories (bilabial, alveolar, palatal, fricative, or other), rather than among 27 individual consonants. Interestingly, the bilabial place of articulation, which is poorly identified in the 'A from A' and 'AV from A' conditions (less than 20 %), becomes quite well identified in the 'AV from AV' condition (75 %). When the bilabial place of articulation is wrongly identified, it is most often mistaken for what we named "non-alveolar fricatives", a group which contains labiodentals or rounded consonants. This suggests that the labial place of articulation remains quite perceptible. Table 6 also shows that the alveolar place of articulation is quite well recovered from the audio alone (approximately 60 %). Although its recovery does benefit from the addition of visual information, the 'AV from AV' score does not reach as high a score as the one for the bilabial place of articulation (70.83 % vs. 75 %). We must recall here that the visual synthesis used in this study does not provide information on tongue movements, which could in fact be visible for alveolar consonants. This probably explains why the addition of visual information does not provide such a drastic improvement on the scores. The palatal place of articulation reaches a good score (above 50

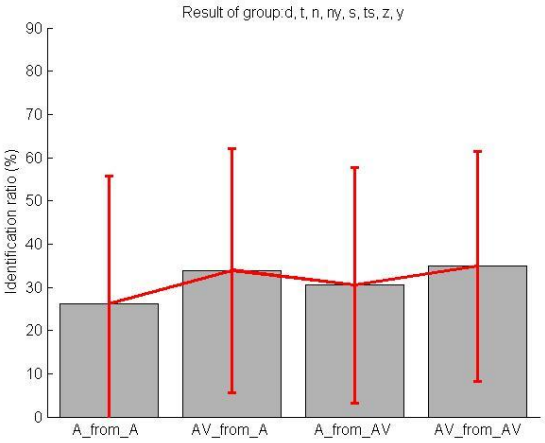
%) but does not benefit much from visual information, as could be expected. The other two groups do not benefit from visual information and are not well identified (less than 50 %).



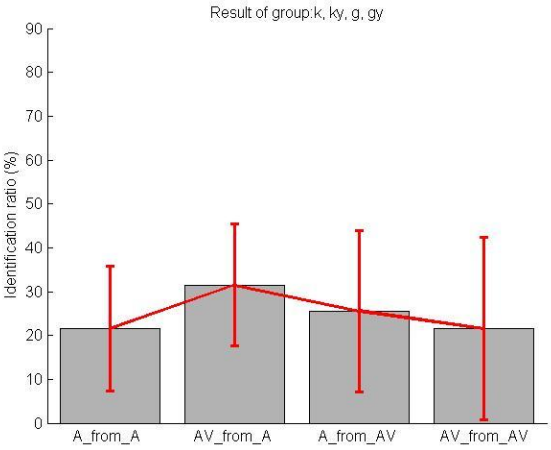
(a) bilabials (F = 2.62 , p < 0.079)



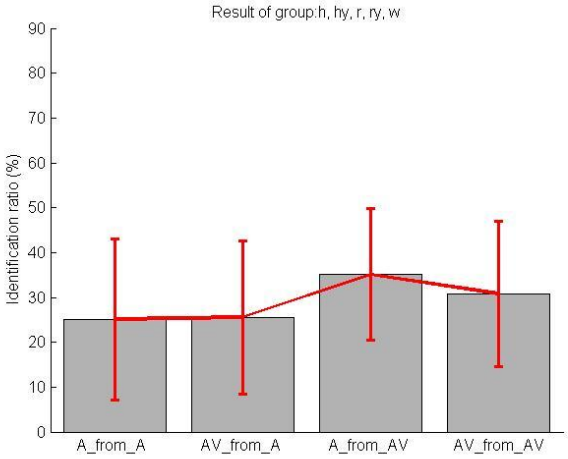
(b) non-alveolar fricatives (F = 0.17, p < 0.912)



(c) alveolars (F = 0.16, p < 0.922)



(d) palatals (F = 0.3 , p < 0.822)



(e) others (F = 0.42, p < 0.744)

Figure 12. Recognition ratios for different groups of consonants, classified according to their places of articulation.

Table 6. Confusion matrices for the 5 places of articulation in the 4 conditions

		bilabial	alveolar	palatal	fricative	other
A from A	bilabial	18.29	10.86	22.86	34.86	13.14
	alveolar	2.38	59.52	11.90	8.73	17.46
	palatal	3.86	15.44	54.83	12.74	13.13
	fricative	0.95	25.71	21.90	47.62	3.81
	other	5.00	20.71	25.71	9.29	39.29
AV from A	bilabial	15.00	10.50	28.00	36.50	10.00
	alveolar	3.47	65.97	9.03	9.72	11.81
	palatal	2.36	16.89	65.20	9.12	6.42
	fricative	0.83	21.67	27.50	49.17	0.83
	other	7.50	19.38	27.50	10.63	35.00
A from AV	bilabial	32.57	5.14	20.57	30.29	11.43
	alveolar	0.79	63.49	15.87	6.35	13.49
	palatal	6.56	13.51	57.14	12.74	10.04
	fricative	3.81	21.90	24.76	47.62	1.90
	other	5.00	20.00	21.43	6.43	47.14
AV from AV	bilabial	75.00	1.00	9.50	10.50	4.00
	alveolar	0.69	70.83	11.81	2.78	13.89
	palatal	1.01	17.91	62.84	10.14	8.11
	fricative	0.83	23.33	26.67	47.50	1.67
	other	4.38	16.88	25.00	6.88	46.88

6. HMM-based whisper-to-speech conversion

6.1. Conversion system overview

In order to compare the performance of the GMM-based voice conversion technique with the approach combining NAM recognition and speech synthesis (Hueber, Chollet *et al.* 2007; Hueber, Chollet *et al.* 2007), an HMM-based whisper-to-speech conversion system was developed. It combines HMM recognition and synthesis: instead of a corpus-based synthesis, we in fact use HMM synthesis (Tokuda, Yoshimura *et al.* 2000). The voice conversion is performed in three steps:

1. Using aligned training utterances, the joint probability densities of source and target parameters are modelled by context-dependent phone-sized HMM. Because of limited training data, we only used the right context, where subsequent phonemes are classified coarsely into 3 groups for vowels ({/a/},{/i/,/e/},{/u/,/o/} without distinguishing between long and short vowels), 2 groups for consonants and 2 groups for utterance-final and -internal silences. Joint observations of each HMM state are modelled with one Gaussian with a diagonal covariance matrix.

2. HMM recognition is performed using the source stream with the HTK toolkit (Young, Kershaw *et al.* 1999). The linguistic model is limited to phone bi-grams learnt on the training corpus.

3. HMM synthesis of the recognized context-dependent phone sequence and target stream is performed using the HTS software (Tokuda, Yoshimura *et al.* 2000; Zen, Nose *et al.* 2007).

6.2. Experiments and results

The same Japanese data (cf. section 5.1) are used: 150 utterances for training and 40 utterances for test. The 0^{th} through 19^{th} mel-cepstral coefficients and their deltas are used as spectral features for both aligned speech and whisper. With a recognition rate of 68.17 % for the input whisper, the spectral distortion between converted speech and target speech is 6.46 dB, compared with 5.99 dB

for the GMM-based system. Note that our recognition rate is quite similar to the 60 % obtained by Hueber *et al.* (Hueber, Chollet *et al.* 2007) using context-independent phones and multimodal input.

The degradation, compared with the GMM-based system, could be explained by the low recognition rate which could be improved by using Gaussian mixtures and more contextual information to expand HMM modelling.

7. Conclusion and perspectives

This paper proposes several solutions to improve the intelligibility and the naturalness of the speech generated by a whisper-to-speech conversion system. The original system proposed by Toda *et al.* is based on a GMM model predicting the three parametric streams of the STRAIGHT speech vocoder (F0, harmonic and noise spectrum) from spectral characterization of the non-audible-murmur input. The first improvement concerns characteristics of the voiced source. We have shown that the estimation of voicing and F_0 by separate predictors improves both predictions. We have also shown that F0 prediction is improved by the use of a large input context window (>400 ms) compared to the original smaller window (90ms). Predictions of all parametric streams are further improved by a data reduction technique that makes use of the phonetic structure of the speech stream.

The second part of the paper compared objective and subjective benefits offered by multimodal data. Improvements obtained by adding visual information in the input stream as well as including articulatory parameters in the output facial animation are very significant.

We will explore several other research directions in the near future: we first plan to use two NAM microphones and source separation techniques to have better signal-to-noise ratios in the input. Source separation can benefit from the joint audiovisual model implicitly captured by the voice conversion system. Such a model can in fact provide very efficient a priori information for source deconvolution techniques (Sodoyer, Girin *et al.* 2004). Finally the voice conversion system can also benefit from other input sources such as ultrasound imaging, surface EMG or EEG

(electroglottography) that can provide additional information on the underlying articulatory speech movements.

A short-term perspective is to develop a portable real-time voice conversion system that will allow us to conduct usage studies and test several experimental variables that are still unexplored, notably the impact of auditory feedback – using the amplified NAM signal or predicted speech - on silent speech production as well as that of conversion delay on production and conversation.

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. We also acknowledge the subjects of the perceptual tests who kindly accepted to identify several hundreds of audio and audiovisual stimuli.

References

[1] Badin, P., G. Bailly, L. Revéret, M. Baciu, C. Segebarth and C. Savariaux (2002). "Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images." Journal of Phonetics 30(3): 533-553.

[2] Bailly, G., A. Bégault, F. Elisei and P. Badin (2008). Speaking with smile or disgust: data and models. AVSP, Tangalooma - Australia, 111-116.

[3] Bailly, G., M. Bérar, F. Elisei and M. Odisio (2003). "Audiovisual speech synthesis." International Journal of Speech Technology 6: 331-346.

[4] Bett, B. J. and C. Jorgensen (2005). Small vocabulary recognition using surface electromyography in an acoustically harsh environment. Moffett Field, CA, NASA, Ames Research Center**:** 16.

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

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

[7] Crevier-Buchman L., Vincent C. & Hans S. (2008). Comportements laryngés en voix chuchotée, étude en caméra ultra-rapide. Actes du 64ième Congrès de la Société Française de Phoniatrie et des Pathologies de la Communication, Paris, octobre 2008.

[8] Heracleous, P., Y. Nakajima, H. Saruwatari and K. Shikano (2005). A tissue-conductive acoustic sensor applied in speech recognition for privacy. International Conference on Smart Objects & Ambient Intelligence, Grenoble - France, 93 - 98.

[9] Higashikawa, M., K. Nakai, A. Sakakura, and H. Takahashi (1996). Perceived pitch of whispered vowels – relationship with formant frequencies: A preliminary study. Journal of Voice, vol 10, No. 2, pp. 155 – 158.

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

[11] Hueber, T., G. Chollet, B. Denby, G. Dreyfus and M. Stone (2007). Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips. Interspeech, Antwerp, Belgium, 658-661.

[12] Hueber, T., G. Chollet, B. Denby, G. Dreyfus and M. Stone (2008). Towards a segmental vocoder driven by ultrasound and optical images of the tongue and lips. Interspeech, Brisbane, Australia

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

[14] Hueber, T., G. Chollet, B. Denby, M. Stone and L. Zouari (2007). Ouisper: Corpus-based synthesis driven by articulatory data. International Congress of Phonetic Sciences, Saarbrücken, Germany, 2193-2196.

[15] Inouye, T. and A. Shimizu (1970). "The electromyographic study of verbal hallucinations." J. Nerv. Mental Disease 151: 415-422.

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

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

[18] Kawahara, H., I. Masuda-Katsuse and A. de Cheveigné (1999). "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.

[19] Nakagiri, M., T. Toda, H. Kashioka and K. Shikano (2006). Improving body transmitted unvoiced speech with statistical voice conversion. InterSpeech, Pittsburgh, PE, 2270-2273.

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

[21] Nakajima, Y. and K. Shikano, "Methods of fitting a nonaudible murmur microphone for daily use and development of urethane elastomer duplex structure type nonaudible murmur microphone", J. Acoust. Soc. Am., 120, 3330 (2006).

[22] Potamianos, G., C. Neti and S. Deligne (2003). Joint audiovisual speech processing for recognition and enhancement. Auditory-Visual Speech Processing, St Jorioz, France, 95-104.

[23] Revéret, L., G. Bailly and P. Badin (2000). 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.

[24] Shimizu, S., M. Otani and T. Hirahara (2009). "Frequency characteristics of several non-audible murmur (NAM) microphones", Acoust. Sci. & Tech. 30, 2, 139-142.

[25] Sodoyer, D., L. Girin, C. Jutten and J.-L. Schwartz (2004). "Developing an audio-visual speech source separation algorithm." Speech Communication 44(1-4): 113-125.

[26] Summerfield, A., A. MacLeod, M. McGrath and M. Brooke (1989). Lips, teeth, and the benefits of lipreading. Handbook of Research on Face Processing. A. W. Young and H. D. Ellis. Amsterdam, Elsevier Science Publishers**:** 223-233.

[27] Summerfield, Q. (1979). "Use of visual information for phonetic perception." Phonetica **36**: 314-331.

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

[29] Toda, T. K. Nakamura, H. Sekimoto and K. Shikano (2009). Voice conversion for various types of body transmitted speech. Proc. ICASSP, pp. 3601-3604, Taipei, Taiwan, Apr. 2009.

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

[31] Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura (2000). Speech parameter generation algorithms for HMM-based speech synthesis. IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 1315–1318.

[32] Tran, V.-A., G. Bailly, H. Lœvenbruck and T. Toda (2008). Predicting F0 and voicing from NAM-captured whispered speech, Proceedings of Speech Prosody, Campinas, Brazil, May 2008.

[33] Walliczek, M., F. Kraft, S.-C. Jou, T. Schultz and A. Waibel (2006). Sub-word unit based non-audible speech recognition using surface electromyography. InterSpeech, Pittsburgh, PE, 1487-1490.

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

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

[36] Zeroual C., Esling J., & Crevier-Buchman L. (2005). Physiological study of whispered speech in Moroccan arabic. Proceedings of Interspeech, Lisbon, 1069–1072.