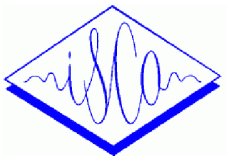
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
(AVSP) 2013
Annecy, France
August 29 - September 1, 2013

Audio-Visual Speaker Conversion using Prosody Features

Adela Barbulescu[1,2], *Thomas Hueber*[1], *Gerard Bailly*[1], *Remi Ronfard*[2]

[1]GIPSA-Lab, CNRS & Universite de Grenoble, St Martin d'Heres, France
[2]IMAGINE team, INRIA / LJK, Grenoble, France

adela.barbulescu@inria.fr, thomas.hueber@gipsa-lab.grenoble-inp.fr
gerard.bailly@gipsa-lab.grenoble-inp.fr, remi.ronfard@inria.fr

Abstract

The article presents a joint audio-video approach towards speaker identity conversion, based on statistical methods originally introduced for voice conversion. Using the experimental data from the 3D BIWI Audiovisual corpus of Affective Communication, mapping functions are built between each two speakers in order to convert speaker-specific features: speech signal and 3D facial expressions. The results obtained by combining audio and visual features are compared to corresponding results from earlier approaches, while outlining the improvements brought by introducing dynamic features and exploiting prosodic features.

Index Terms: speaker identity conversion, gaussian mixture model, dynamic features, prosodic features

1. Introduction

Speaker identity conversion refers to the challenging problem of converting multimodal features between different speakers such that the converted performance of a source speaker can be perceived as belonging to the target speaker. One representative feature for speaker individuality is the speaker's voice, and the area of speech processing related to the topic is voice conversion. Moreover, human speech presents a bimodal nature, as speech perception is also influenced by visual cues, represented by lip movements or facial expressions. The audio-visual interaction infers a high significance in human speech perception. Due to the various applications of multimodal interaction, a large amount of research has been conducted in this area and the interest has also been directed towards joint audio-visual processing [1].

This paper addresses the problem of speaker conversion using audio and 3D visual information, $i.e.$ the speech signal and the 3D scans of a source speaker for a certain utterance will be modified to sound and look as if uttered by a target speaker. Applications of the presented work are found in various fields, with specific examples such as audio-visual puppetry, facial animation retargetting [2] or automated 3D animation.

Although the speech community has showed increased attention in voice conversion, it remains a challenging problem, one reason being the subjectivity of perceived conversion quality. Also, the difficulty in processing the speech characteristics which confer voice individuality affects the quality of voice conversion. Such characteristics may be of linguistic or non-linguistic type. The non-linguistic characteristics proved to have a high influence on speaker individuality and can be classified as sociological (social class, region of birth, age) and physiological factors (shape of the vocal tract). The former affect the speaking style which may be described by prosodic features such as pitch contour, speaking rate, duration of words and pauses etc. The latter play a high role in individual voice quality and are strongly connected to the spectral content [3].

Most existing voice conversion systems use only spectral envelope characteristics and prosodic features, such as pitch frequency features and overall speech dynamics. A great deal of speaker-specific information can be carried by averaged values of the prosodic features [4] [5], this being also a reason for which finer prosodic features are ignored in most VC systems. Emotional voice conversion using prosodic features has been referred in a few late works [6] [7].

Our paper also uses spectral conversion by building statistical relations between the spectral envelopes of a source and target speaker. Thus, by training on experimental data consisting of identical utterances given by two speakers a mapping function is built to convert the source spectral features. We introduce a prosodic feature to describe the speaker-specific rhythm of speech by exploiting the alignment between the two speech signals.

The fusion of audio video information has been used in improving speech recognition systems [8] and also in a few works on speaker identification [9] [10]. The visual information is usually provided by processing lip texture and lip motion and the extracted features are needed in speaker discrimination [11]. On the other hand, multimodal speaker conversion has been addressed only in a few works which include late integration of audio video information $i.e.$ speech and video data are processed separately and recombined at the end. Similar to our paper is the work presented in [12] where speaker conversion rep-

resents the late fusion of converted speech and 3D facial movements.

The paper is organized as follows: section 2 presents theoretical aspects regarding the approach used for feature extraction and conversion. Section 3 describes the experimental dataset and implementation details. Next, the results and experimental evaluation of the methods presented are described in section 4, while conclusions and future perspectives are presented in section 5.

2. Feature conversion approaches

Most voice conversion systems use statistical approaches to create the feature conversion function from experimental data. The state of the art performs spectral conversion using Gaussian mixture model and a similar approach is used in our paper for joint audio-video features.

2.1. Gaussian Mixture Model conversion

The basic conversion approach implies modeling data using GMMs and building a conversion function as a weighted sum of local regression functions, thus providing a soft classification between mixture components [13] [14]. In the following example, the joint probability density of source and target feature data represented by parameter vectors $x_t = [x_t(1), x_t(2), ..., x_t(D_x),]^T$ and $y_t = [y_t(1), y_t(2), ..., y_t(D_y)]^T$ at frame t, are modeled by the GMM:

$$P(z_t|\lambda^{(z)}) = \sum_{m=1}^{M} \alpha_M \mathcal{N}(z_t; \mu_m^{(z)}, \Sigma_m^{(z)}), \qquad (1)$$

where z_t is the joint vector $z_t = [x_t^T, y_t^T]^T$, m is the mixture component index corresponding to the weight α_m. $\mathcal{N}(\cdot; \mu, \Sigma)$ represents a normal distribution with mean μ and covariance Σ, defined as follows:

$$\mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}, \qquad (2)$$

Given a source vector x and m being the GMM mixture component index, the conditional probability density of the converted vector y is also modeled as a GMM with the mean and covariance:

$$E_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} {\Sigma_m^{(xx)}}^{-1} (x_t - \mu_m^{(x)}), \qquad (3)$$

$$D_m^{(y)} = \Sigma_m^{(yy)} - \Sigma_m^{(yx)} {\Sigma_m^{(xx)}}^{-1} \Sigma_m^{(xy)}. \qquad (4)$$

The minimum mean square estimation of the converted vector is:

$$\hat{y}_t = E[y_t|x_t] = \sum_{m=1}^{M} w_m E_{m,t}^{(y)}, \qquad (5)$$

where the weight w_m represents posterior probability of x for the mth component:

$$w_m = P(m|x_t, \lambda^{(z)}) = \frac{\alpha_m \mathcal{N}(x_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(x_t; \mu_x^{(x)}, \Sigma_n^{(xx)})}, \qquad (6)$$

Despite the popularity and good mapping functionality of the conventional method, the results can be improved by solving the problems created by time-independent mapping and oversmoothing. Time-independency assumes that each vector is converted on a frame-basis, disregarding the information contained in other frames. Therefore, the converted vector suffers from discontinuities and a solution is represented by introducing the dynamic features of vector parameters.

Toda [15] introduces the maximum likelihood estimation of parameter trajectory in which feature vectors are converted simultaneously over a time sequence. The source and target features at frame t consist of static and dynamic features: $X_t = [x_t^T, \Delta x_t^T]$ and $Y_t = [y_t^T, \Delta y_t^T]$ and the parameter vectors over an utterance are regarded as a single time-sequence vector: $X = [X_1^T, X_2^T, ..., X_T^T]$ and $Y = [Y_1^T, Y_2^T, ..., Y_T^T]$. The target time-sequence vector including dynamic features is computed from the initial static target vector: $Y = Wy$, where W is a $[2D_yT] - by - [D_yT]$ matrix of predefined weights.

Given a source vector X and the parameter set $\lambda^{(Z)}$ of the GMM trained on the joint vector $Z = [X^T, Y^T]$, the MLE-based mapping determines the converted target vector as follows:

$$\hat{y}_t = arg\,max\, P(Y|X, \lambda_Z) \qquad (7)$$

The maximization in (7) is done by maximizing an auxiliary function such that the converted vector sequence is given for the suboptimal approximation $\hat{m} = arg\,max\, P(m|X, \lambda_Z)$ by:

$$\hat{y} = (W^T {D_{\hat{m}}^{(Y)}}^{-1} W)^{-1} W^T {D_{\hat{m}}^{(Y)}}^{-1} E_{\hat{m}}^{(Y)}, \qquad (8)$$

where

$$E_{\hat{m}}^{(Y)} = [E_{\hat{m}_1,1}^{(Y)}, E_{\hat{m}_2,2}^{(Y)}, ..., E_{\hat{m}_T,T}^{(Y)}], \qquad (9)$$

$${D_{\hat{m}}^{(Y)}}^{-1} = diag[{D_{\hat{m}_1}^{(Y)}}^{-1}, {D_{\hat{m}_2}^{(Y)}}^{-1}, ..., {D_{\hat{m}_T}^{(Y)}}^{-1}]. \qquad (10)$$

Unlike in the MMSE method, the converted target vector is computed as a weighted sum of the mean vectors where covariance matrices are used as weights. The covariance matrices can be regarded as a confidence measure of conditional mean vectors from individual mixture components.

2.2. Prosodic feature conversion

A mapping function between source and target speech signals can be built if the two signals are aligned according to a chosen criteria. In our case, an alignment

path between corresponding frames is retrieved using a Dynamic Time Warping algorithm. A correlation is observed between the local slope of the path and the rhythm of speech. Thus, by computing local slope on a small window-frame we extract a new prosodic feature that is speaker-specific. At frame t, for the path alignment between two speakers, the slope parameter is given by:

$$slope_t = \frac{\Delta p_y}{\Delta p_x} \tag{11}$$

where p_y and p_x represent the speaker-specific alignment path vectors. The conventional MMSE approach (5) can be applied by concatenating the slope parameter to the spectral feature vector in order to estimate the speaker-specific rhythm of speech for a new utterance.

3. Implementation and feature description

3.1. 3D audio-video database

Experiments were conducted on the Biwi 3D Audiovisual Corpus of Affective Communication [16] comprising a total of 1109 sentences (4,67 seconds long on average) uttered by 14 native English speakers (6 males and 8 females). The dense dynamic face scans were acquired at 25 frames per second by a realtime 3D scanner and the voice signal was captured by a professional microphone at a sampling rate of 16kHz. Along with the detailed 3D geometry and texture of the performances, sequences of 3D meshes are provided, with full spatial and temporal correspondences across all sequences and speakers. For each speaker 80 utterances are recorded, half in a personal speaking style and half in an "emotional" manner, as they are asked to imitate an original version of the performance.

3.2. Feature extraction

Spectral features are extracted at a segmental level using the STRAIGHT vocoder [17] which decomposes speech into a spectral envelope without periodic interferences, F_0 and relative voice aperiodicity. These parameters are further encoded and from the spectral envelope, we use the 1st through 24th Mel-cepstral coefficients, a widely made choice in VC and voice synthesis/analysis systems [13] [15].

The speaker-specific facial articulation features are captured from a dense mesh of 3D data. From the dataset, 7 speech and expressive movement components are extracted following a guided Principal Component Analysis method [18] [19]. As the mouth opening and closing movements have a large influence on face shape, the first component is used as a first predictor, iterative PCA is performed on residual lips values and the next 3 lips components are obtained. The second jaw component is used as the 5th predictor and the last two parameters are extracted as expressive components and represent the zygotic and eyebrow muscle movements. These features are computed at the original video frame rate and are later oversampled to match the audio frame rate. Both visual and spectral features are concatenated with their first derivatives in order to be used for the MLE-based mapping approach described in the previous section.

Speaking style features are extracted by computing the local alignment path slope on a 6-dimensional frame window for each frame and the slope parameter is concatenated to the spectral feature vector. Frame sampling is done according to the predicted local slope parameter such that the total duration of the converted signal and local "rhythm" of speaking are varied.

4. Experimental evaluation

Experimental evaluations are conducted with the following goals: to assess the quality of the converted signals, the effectiveness of dynamic features and to compare results between joint and separate audio-visual training. All possible conversions are performed between speakers of different age, race and sex, using 40 sentences for training and 40 for testing per speaker. Four combinations of input features are used in separate experiments: spectral features, spectral features and slope parameter, video features, combined spectral and video features. The optimal number of mixture components chosen considering the size of the training dataset and feature vectors is 16.

4.1. Objective evaluation

4.1.1. Evaluation on audio and visual features

Objective evaluation of spectral conversion is done by measuring the cepstral-distortion between target signal and interpolated converted signal using the following equation:

$$Mel - CD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(mc_d^{(y)} - \hat{mc}_d^{(y)})^2} \tag{12}$$

where $mc_d^{(y)}$ and $\hat{mc}_d^{(y)}$ represent the d-th spectral coefficient of the target and converted signal respectively. The error measure for 3D facial expressions conversion is obtained by reprojecting the converted visual parameters to 3D point coordinates and computing the distance between target and converted 3D points:

$$error[cm] = \sqrt{\sum_{d=1}^{7}(p_d^{(y)} - \hat{p}_d^{(y)})^2} \tag{13}$$

where $p_d^{(y)}$ and $\hat{p}_d^{(y)}$ represent the d-th visual parameter of the target and converted signal respectively.

Table 1 presents the original and converted values obtained for mel-cepstral distortion and distance error for

all combinations of feature vectors used in the mapping process: from audio to audio, from visual to visual, from audio-visual to audio-visual, from audio-visual to audio and from audio-visual to visual. Evaluation on joint and

Table 1: *The first three columns present mel-cepstral distortions (dB) for audio feature evaluation and the last three present distance error (cm) for video feature evaluation. Results are obtained from converting all speakers to M6.*

input	A	AV	AV	V	AV	AV
output	A	A	AV	V	V	AV
original	6.86	6.86	6.86	0.75	0.75	0.75
MMSE	5.30	5.80	5.80	0.37	0.38	0.38
MLE	5.05	5.47	5.47	0.34	0.34	0.33

separate feature training is done by comparing results obtained on conversions between all speakers and a male. In all feature combinations used, the MLE approach retrieves a smaller prediction error than in the conventional approach, due to the use of additional inter-frame information. Moreover, the use of dynamic features for visual conversion generates perceptually improved results as discontinuities between frames are smoothed.

The converted signal obtained by joint training tends to follow the target signal patterns better than in the case of separate audio or video training. The type and size of feature vectors used have an influence on the obtained results; in the case of audio conversion where the smallest errors are obtained for identical input-output feature combinations, with a close distortion rate both for separate and joint audio-visual features. Figures 1 and 2 present the mean error values obtained for the presented conversion approaches for all speaker pairs. The original distortion and distance error values are closer for the male speakers (M1 to M6). The higher difference between original and predicted distance error for visual features can be explained by the morphological differences between source and target speakers.

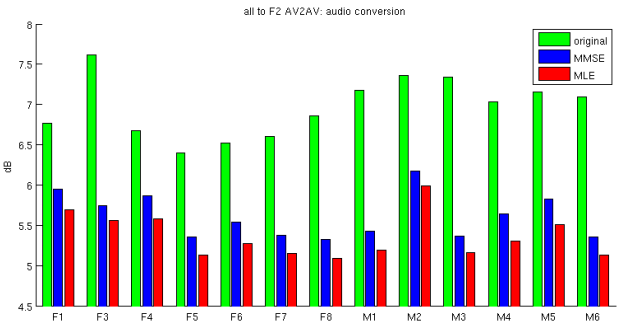


Figure 1: Mel-cepstral distortion values for joint feature conversion between all speakers and female speaker F2.

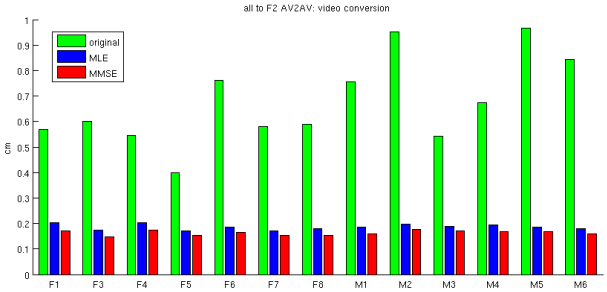


Figure 2: Distance error values for joint feature conversion between all speakers and female speaker F2.

As shown in other works on VC systems, distortion computation is not necessarily correlated with perceptual results therefore a subjective evaluation is needed to assess the performance of our system. The performance of spectral conversion is measured by synthesizing converted speech signal with the STRAIGHT framework, using an averaged F_0 between source and target. The converted speech signal and 3D expression parameters are used with the speaker rigid scan information in order to generate animations of the performance. The animations can further be used for subjective evaluations regarding speaker individuality.

4.1.2. Evaluation on slope prosodic feature

One evaluation method for the slope parameter is assessing the correlation coefficients computed from a matrix M that is composed of original and predicted slope parameters:

$$R_{i,j} = \frac{C_{i,j}}{\sqrt{C_{i,i}C_{j,j}}} \tag{14}$$

where C is the covariance matrix of M.

Figure 3 presents results obtained from converting the speech signal including the slope parameter from all speakers to one male speaker.

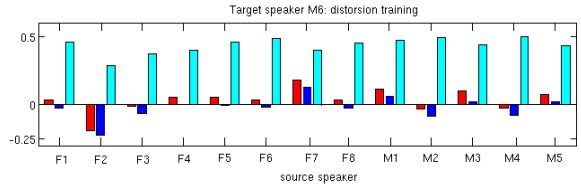


Figure 3: Conversion from all speakers to M6. The red bars represent original slope values, the blue bars predicted slope and cyan bars the correlation between original and predicted.

Strong dissimilarities can be observed for source speakers F2 and F7, which have slower and faster rhythms of speech, respectively. A distribution of orig-

inal and predicted slope values according to associated phonemes is presented in figure 4.

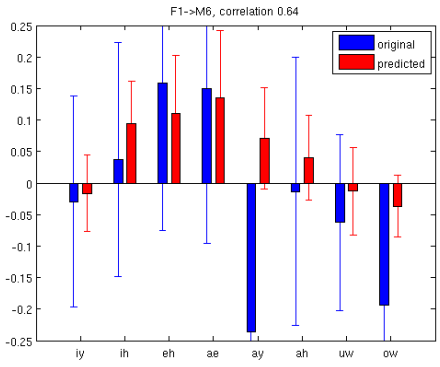


Figure 4: Original and predicted slope values (blue and red bars respectively) according to analyzed segments by conversion from F1 to M6.

We are interested in partitioning in a restricted set of acoustic zones, revealing the different slope behaviours on closed and open vowels and thus showing that the articulation rate depends also on phonetic content. Figure 5 illustrates the original and predicted slope parameters values computed for all frames during one utterance. The predicted values follow the local patterns of the original slope thus predicting the speaker-specific rhythm of speech.

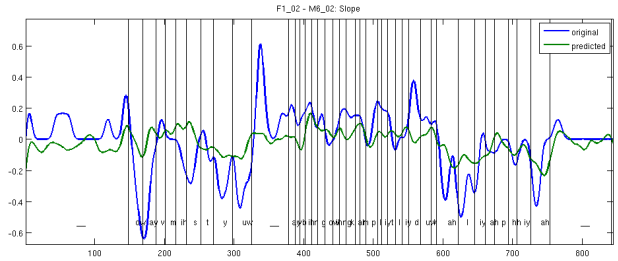


Figure 5: Original and predicted slope values for corresponding phonemes of the utterance converted from F1 to M6.

4.2. Subjective evaluation

Speaker individuality evaluation is conducted using a category change test. The listener is trained with natural speech belonging to target speaker M6 and is then tested on a set of randomized utterances. For each utterance the listener presses a certain key if the speaker is an "impostor". The test set includes groups of 4 types of utterances representing the same sentence: original target, simple converted speech, converted speech with slope feature and interpolated source along DTW path with averaged F_0. The converted speech should be perceived as belonging to the target speaker while the interpolated source signals represents impostor speakers. The test is reduced to a Yes-No task as the listener states whether an utterance belongs to speaker M6 or not. The test set is composed of 20 utterances and the number of listeners is 13. Figure 6 shows the percentage of impostors chosen from each group of utterances defined by the generation method. As expected, the actual impostors are tracked with the highest percentage (98%) and there is a 90% difference in detections between the impostor group and the converted and natural sequences groups. Moreover, there are more impostor detections in the group obtained by simple conversion than in the one obtained by conversion with slope.

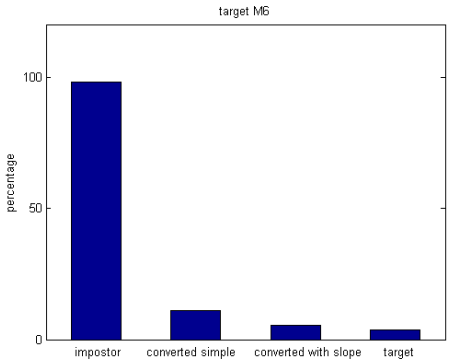


Figure 6: Percentage of impostors chosen per types of sequences.

5. Conclusion and perspectives

The work presents an approach towards speaker identity conversion using speech signals and 3D facial expressions. The GMM-based method using dynamic features which was introduced for VC systems is used here for different types of input features: spectral, video parameters, joint audio-video. Moreover, prosodic features are extracted from time alignment information for a better conversion of speaking styles. Objective experimental results on different combinations of speaker conversions from the 3D BIWI dataset have demonstrated the effectiveness of dynamic features and the subjective evaluation illustrates that the converted sequences are perceived as belonging to target speaker. However a more extended dataset and experiments are needed to assess the introduction of the prosodic feature and the effectiveness of using a joint training dataset.

6. References

[1] Chen, T. and Rao, R., "Audio-visual integration in multimodal communication", Proceedings of IEEE, Special Issue on Multimedia Signal Processing, pp. 837-852, 1998.

[2] Elina, H., Gabbouj, M., Nurminen, J., Siln, H. and Popa, V.," Speech enhancement, modeling and recognition algorithms and applications", 2012.

[3] Weise, T. et al,"Face/off: Live facial puppetry", Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, ACM, 2009.

[4] Helander, E. and Nurminen, J., "On the importance of pure prosody in the perception of speaker identity", Proc. of Interspeech, 2007.

[5] Kuwabara, H. and Sagisak, H., "Acoustic characteristics of speaker individuality: Control and conversion", Speech communication 16.2: 165-173, 1995.

[6] Aihara, R. et al."GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features", American Journal of Signal Processing 2.5: 134-138, 2012.

[7] Tao, J., Kang, Y. and Li, A., "Prosody conversion from neutral speech to emotional speech", Audio, Speech, and Language Processing, IEEE Transactions on 14.4: 1145-1154, 2006.

[8] Petajan, E. D., "Automatic lipreading to enhance speech recognition", Proc. IEEE Global Telecommunication Conf., Atlanta, 1984.

[9] Frischholz, R., Dieckmann, U., "BioID: a multimodal biometric identication system", J. IEEE Comput. 33 (2) 6468, 2000.

[10] Erzin, E., Yemez, Y., Tekalp, A., "Multimodal speaker identication using an adaptive classier cascade based on modality reliability", IEEE Trans. Multimedia 7 (5) 840852, 2005.

[11] etingl, H. E. et al. "Multimodal speaker/speech recognition using lip motion, lip texture and audio", Signal processing 86.12: 3549-3558, 2006.

[12] Inanoglu, Z., Jottrand, M., Markaki, M., Stankovic, K., Zara, A., Arslan, L., Dutoit, T., Panzic, I., Saralar, M. and Stylianou, Y., "Multimodal Speaker Identity Conversion", eNTERFACE, 2007.

[13] Stylianou, Y., Capp, O. and Moulines, E., "Continuous probabilistic transform for voice conversion", Speech and Audio Processing, IEEE Transactions on 6.2: 131-142, 1998.

[14] Kain, A., and Macon, M. W., "Spectral voice conversion for text-to-speech synthesis", Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on. Vol. 1. IEEE, 1998.

[15] Toda, T., Black, A. W.and Tokuda, K., "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory", Audio, Speech, and Language Processing, IEEE Transactions on 15.8: 2222-2235, 2007.

[16] Fanelli, G., Gall, J., Romsdorfer, H., Weise, T. and Van Gool, L.,"Acquisition of a 3d audio-visual corpus of affective speech", IEEE Transactions on Multimedia, Vol. 12, No. 6, pp. 591 - 598, 2010.

[17] Kawahara, H., "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds", Acoustical science and technology 27.6: 349-353, 2006.

[18] Bailly, G. et al, "Lip-synching using speaker-specific articulation, shape and appearance models", EURASIP Journal on Audio, Speech, and Music Processing 2009: 5, 2009.

[19] Revret, L., Bailly, G. and Badin, P., "MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation", 2000.