GMM Mapping Of Visual Features of Cued Speech From Speech Spectral Features

Zuheng Ming, Denis Beautemps, Gang Feng

GMM Mapping Of Visual Features of Cued Speech From Speech Spectral Features

Zuheng Ming[1,2], *Denis Beautemps*[1,2], *Gang Feng*[1,2]

[1] Univ. Grenoble Alpes, GIPSA-lab, F-38402 Saint Martin d'Hères Cedex
[2] CNRS, UMR 5216

Denis.Beautemps@gipsa-lab.grenoble-inp.fr

Abstract

In this paper, we present a statistical method based on GMM modeling to map the acoustic speech spectral features to visual features of Cued Speech in the regression criterion of Minimum Mean-Square Error (MMSE) in a low signal level which is innovative and different with the classic text-to-visual approach. Two different training methods for GMM, namely Expecting-Maximization (EM) approach and supervised training method were discussed respectively. In comparison with the GMM based mapping modeling we first present the results with the use of a Multiple-Linear Regression (MLR) model also at the low signal level and study the limitation of the approach. The experimental results demonstrate that the GMM based mapping method can significantly improve the mapping performance compared with the MLR mapping model especially in the sense of the weak linear correlation between the target and the predictor such as the hand positions of Cued Speech and the acoustic speech spectral features.

Index Terms: Cued Speech, LSP, MFCC, GMM mapping.

1. Introduction

The framework of this paper is speech communication for deaf orally-educated people. Speech is concerned here in its multimodal dimensions and in the context of automatic processing. Indeed, the benefit of visual information for speech perception (called "lip-reading") is widely admitted. However, even with high lip reading performances, without knowledge about the semantic context, speech cannot be thoroughly perceived. The best lip readers scarcely reach perfection. On average, only 40 to 60% of the phonemes of a given language are recognized by lip reading ([1]), and 32% when relating to low predicted words ([2]) with the best results obtained amongst deaf participants - 43.6% for the average accuracy and 17.5% for standard deviation with regards to words ([3], [4]), with an advantage of deaf women with 33.3 % for females but only 23.5 % for males (see the study on word perception of [5]). The main reason for this lies in the ambiguity of the visual pattern. However, as far as the orally educated deaf people are concerned, the act of lip-reading remains the main modality of perceiving speech. This led Cornett ([6]) to develop the Cued Speech system (CS) as a complement to lip information. CS is a visual communication system that makes use of hand shapes placed in different positions near the face in combination with the natural speech lip-reading to enhance speech perception from visual input. This is a system (See Figure 1) where the speaker, facing the perceiver, moves his hand in close relation with speech (See [7] for a detailed study on CS temporal organization in French language).

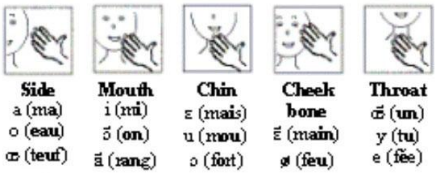

Figure 1: *Hand placements for coding vowels in French Cued Speech (from [7]).*

CS is largely improving speech perception for deaf people ([2]), relating to the identification of American-English syllables; and ([8]), relating to the identification of sentences in American-English language (scores between 78 and 97%). Moreover, CS offers to deaf people a thorough representation of the phonological system, inasmuch as they have been exposed to this method since their youth, and therefore it has a positive impact on the language development (see [9] for CS in French language).

As we have seen from this short review, the Cued Speech method offers a real advantage for complete speech perception. Nowadays, one of the important challenges is the question of speech communication between normal hearing people who do not practice CS but produce acoustic speech and deaf people with no auditory rests who use lip-reading completed by CS code for speech perception. To solve this question, one can use a human translator. Another solution is based on the development of automatic translation systems. This paper is a contribution to this topic. In a more general framework, two sources of information could contribute to this translation operation: (i) The *a priori* knowledge of the phonetic, phonologic and linguistic constraints; (ii) the *a priori* knowledge of the correlations between the different vocal activity: neuronal and neuro-muscular activities, articulatory movements, aerodynamic parameters, vocal tract geometry, face deformation and acoustic sound. Different methods allow their modelling and their optimal merging with the input signals and the output ones. On an axis ordering the methods in function of their dependence upon the used language, one can find at the two extremities: (i) the method using the phonetic level of interface, combining speech recognition and speech synthesis to take into account speech phonology organization. Note that the recognition and synthesis processing can call on very various modelling techniques. If the phonetic models based on Hidden Markov Models (HMM) are the basis of the main recognition systems, synthesis based on concatenative of multi-parametered units of various lengths is still very popular. Note the increasing interest of synthesis by trajectory models based on HMMs ([10], [11]) allowing the jointed learning of the recognition and synthesis systems; (ii) The methods using the correlation between signals without the help of the phonetic level but using various mapping techniques. These techniques capture the correlations between

input and output samples using Vector Quantification or Gaussian Mixture Model ([12], [13]). For Cued Speech, the classic method to convert audio speech to CS components consists of coupling a recognition system to a text-to-visual speech synthesizer ([7], [14], [15], [16]). The link between the two systems requires at least the phonetic high level. Before this work, no studies aimed at using the very low signal level. This work is a contribution to this challenge in the case of oral French vowels. A new approach based on the mapping of speech spectral parameters with the visual components made of CS and lip parameters is proposed. In this context the objective of the mapping process is to deliver visual parameters that can be used as target parameters for visual speech synthesis. In this paper, we explore the GMM-based mapping. Berthommier ([17]) applied a similar procedure to estimate the DCT coefficients of the lip region in the objective of speech enhancement of noisy signal. The present paper deals with normal speech in the case of speech supplemented by CS. In the following, we will start by defining the audio and visual parameters which will be taken into account in the mapping process. Then we will first present the results with the linear approach and then the improvement obtained with the use of multiple GMMs.

2. Experimental set-up and spectral, lip and Cued Speech material

2.1. Database recording

The data have been derived from a video recording of a speaker pronouncing and coding in CS a set of 50 isolated French words. The words were made of 32 digits (from 0 to 31), 12 months and 6 more ordinary words. Each word was presented once on a monitor placed in front of the speaker, in a random order. The corpus has been uttered 10 times. The speaker is a female native speaker of French graduated in CS. The recording has been made in a sound-proof booth and the image video recording rate was set on 25 image/second. The speaker was seated in front of a microphone and a camera connected to a Betacam recorder. Landmarks were placed between eyebrows and at the extremity of the fingers to further extraction of the coordinates used as Cued Speech hand parameters. In addition, a square paper was recorded for pixel-to-centimeter conversion.

The video recording has been done with the PAL format, thus saved as numerical Bitmap RGB images made of the interlaced half-frames of the video (respectively even and odd lines). Each image was de-interlaced into two half-frames and the missing lines of the each half-frame were filled by linear interpolation, as to obtain two de-interlaced full frames corresponding to two recordings separated with 20 ms.

2.2. Extraction of lip and hand visual features

These frames constitute the set of images at the rate of 50 Hz that we will refer to in the following. For its part, the audio of the recording was digitalized at 44100 Hz and re-sampled at 16000 Hz. For each word, the coordinates of the inner contour of the lips have been manually selected on the corresponding images and converted into centimeters with the use of the pixel-to-centimeter conversion equation. Finally, the following geometric lip features were derived (following [18]): the lip width (A), the lip aperture (B) and the lip area (S) respectively. The work presented in this paper focuses on vowels. The database has thus been made by the vowels extracted from this set of material. The audio signal was used as to first locate the vowels inside the isolated words, then to derive the corresponding video frames. Thereafter the t_0 instants in which the lips were at the corresponding target were precisely defined from the analysis of the subset of video frames. 16 LSP coefficients were derived from the audio on the basis of a 20 ms Hamming window centered on t_0 together with 16 MFCC coefficients calculated on the basis of a 32 ms Hamming window. In addition 4 formants were derived from the spectral envelop (obtained with the LSP coefficients). Since the Cued Speech hand position target are often not synchronous with variation of lips or speech, the t_1 instants for Cued Speech hand target are selected separately by analysis of the subset of video frames. Then the hand features defined as the relative (x,y) coordinates of the fingertip of the middle finger (or index finger if middle finger is missing) in reference to the landmark between eyebrows were extracted. The whole of these processing thus made it possible to constitute a database made of 1371occurrences of the 10 French vowels (table 1). In the next, the accuracy of the mapping methods will be measured using a 1/5 cross-validation test. For that, the 1371 occurrences were divided into 5 partitions made of approximate 275 elements for each partition. Finally, 4 principal components (derived from a PCA Analysis) of the 4 formants, 16 principal components of the LSPs, 16 principal components of the MFCCs, the totally 32 principal components of the set of the LSP and MFCC coefficients, the (x, y) coordinates of the hand and the (A, B, S) lip parameters were derived for each element of each of the 5 partitions.

Table 1. *List of the ten French vowels with their occurrence*

Vowels	[i]	[e]	[ɛ]	[a]	[y]	[ø]	[œ]	[ɔ]	[o]	[u]
Occurrence	236	255	231	168	37	80	137	83	40	104

3. The Multiple-linear regression based mapping modeling

In this section, the objective is to predict the (A, B, S) lip parameters and the (x,y) hand coordinates with the corresponding spectral parameters (their principal components F_i) using the multiple-linear modeling.

3.1. The used method

The set of F_i was ordered in function of their "prediction power" using their ρ correlation coefficient with the parameter to be predicted. The F_i predictors were then sorted following the decreasing values of their ρ^2 as to obtain the sorted $F = [F_1, F_2, ..., F_p]$, $(1 \leq p \leq 32)$. In the following, the method is illustrated with the lip parameter B defined as the target, after being centered. B was submitted to a linear regression with the first predictor F_1. The linear coefficient k_1 was obtained as to minimize the residual error between the real values of the target and the predicted ones in the sense of least square error. The residual error was then submitted to a linear regression with the second predictor F_2 and so on until the p order. Finally, the estimation equation at the order p is the following:

$$\widehat{B} = k_1 F_1 + k_2 F_2 + ... + k_p F_p + \overline{B} \qquad (1)$$

$$\boldsymbol{k} = (F^T F)^{-1} F^T (B - \overline{B}) \quad (2)$$

Where, $\boldsymbol{k} = [k_1, k_2, ..., k_p]^T$. As mentioned before, the 5 partitions of the database was used to evaluate the accuracy of the mapping. One of the partitions was reserved for testing by turns, while the other 4 partitions were used for the training by applying the estimation equation (1). The residual variance as the complement of the explained variance of the considered lip parameter was calculated. Finally, the average residual variance was calculated over the 5 combinations of the training and testing partitions for evaluating the model.

3.2. Results for the lip parameters

In the following, the average residual variance calculated over the 5 combinations of the training partitions is considered. Figure 2 plots the average residual variance of lip parameter B in function of the number of predictors. From the figure, it can be first observed that the residual variance decreases in function of the number of used predictors. One can then notice that the residual variance remains high with the use of formants (around 39 % of the initial variance). This is probably due to a lack of dimensions. Indeed the 16 MFCC and LSP coefficients improve very significantly the performances of the prediction (the residual variances are 25% and 18% respectively). The MFCCs allow a quicker decrease while the LSP coefficients attain a lower residual variance. Finally, the prediction based on the mixture of the MFCC and LSP has the advantage of the quick decrease property of the MFCCs and the low residual of the LSP. This mixture of MFCCs and LSP is thus considered as the best parameters for this prediction even if the final error is still relatively high (around 14 % on the training database). The prediction performance of the other two lip parameters A and S are similar to situation of the B. These results will be used as a reference for the following, in particular for the choice of the set of pertinent predictors. Finally, we obtained very similar results with the test data.

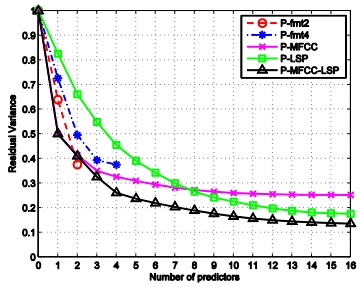

Figure 2: *The average residual variance of the lip parameters B over the 5 combinations of the training partitions, in function of the number of predictors (based on 2 formants (P-fmt2), 4 formants (P-fmt4), MFCC, LSP and the mixture of MFCC-LSP).*

3.3. Results for the hand parameters

The same method of analysis was applied for predicting the (x,y) coordinates of the hand. The final value of the residual variance reaches 39 % for x and 29 % for y, even in the case of the best predictors made up of the whole of the LSP parameters and MFCCs (see Figure 3). This high value of the final residual variance is explained by the low values of the correlations coefficient between the predictors and the target (0.43 and 0.42 respectively with x and y). This weak linear correlation between the spectral parameters and the (x, y) coordinates of the hand positions probably gives rise to the limit of the linear method.

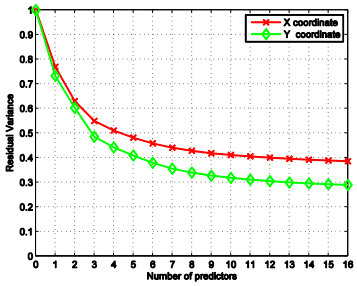

Figure 3*: The average residual variance of the hand coordinates on the 5 combinations of the training partitions with the first best predictors of the whole set of MFCC-LSP in function of the number of predictors.*

In order to check this assumption, the Cued Speech (x, y) coordinates have been re-organized in a coherent way with the French vocalic triangle defined by the formant space consisted of the first two formants, given the strong linear correlation between formants and the spectral parameters LSP (The maximum linear correlation coefficients between the spectral parameters LSP and the first two formants are 0.96 and 0.87 respectively). Therefore a great fall of the residual variances could be obtained which finally reach 7.85% and 7.08% respectively for the redistributed x and y coordinates.

4. The GMM based mapping model

4.1. The Method

In this section, the principal components of the set of 16 LSP and 16 MFCC coefficients constitute the source vector $\mathbf{x}$ with dimension p ($1 \leq p \leq 32$) and the lip or hand parameters are the target vector $\mathbf{y}$. In reference to the equation (3) (see also [19], [20]), the estimator (in the sense of MMSE) of the parameter $\mathbf{y}$ has a linear regression form of observation $\mathbf{x}$ weighted by the a posteriori conditional probability of component c_i:

$$\widehat{\mathbf{y}} = F(\mathbf{x}) = \sum_{i=1}^{m} (W_i \mathbf{x} + b_i) P(c_i|\mathbf{x}) \quad (3)$$

Where, $P(c_i|\mathbf{x})$ is the a posteriori conditional probability that observation $\mathbf{x}$ is generated by the c_i component/Gaussian with the mean vector $\boldsymbol{\mu}_i^{\mathrm{x}}$ and covariance matrix $\boldsymbol{\Sigma}_i^{\mathrm{XX}}$, W_i and b_i being the transform and the bias matrices respectively associated to component c_i.

$$b_i = \boldsymbol{\mu}_i^{\mathrm{Y}} - \boldsymbol{\Sigma}_i^{\mathrm{YX}} (\boldsymbol{\Sigma}_i^{\mathrm{XX}})^{-1} \boldsymbol{\mu}_i^{\mathrm{X}} \quad (4)$$

$$W_i = \boldsymbol{\Sigma}_i^{\mathrm{YX}} (\boldsymbol{\Sigma}_i^{\mathrm{XX}})^{-1} \quad (5)$$

$$P(c_i|\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i^{\mathrm{X}}, \boldsymbol{\Sigma}_i^{\mathrm{XX}})}{\sum_{j=1}^{m} \alpha_j N(\mathbf{x}; \boldsymbol{\mu}_j^{\mathbf{x}}, \boldsymbol{\Sigma}_j^{\mathrm{XX}})} \quad (6)$$

Where, α_i is the weighting coefficient of the Gaussian model, the sum of all the coefficients is 1; $\boldsymbol{\Sigma}_i^{\mathrm{YX}}$ is the matrix of covariance between $\mathbf{x}$ and $\mathbf{y}$ and calculated on the component

c_i ($1 \leq i \leq m$) subset of data and $\boldsymbol{\mu}_i^{\mathrm{Y}}$ is the mean vector of the target vector on this same subset. Note that when the number of the Gaussian equals to one, namely m=1, the GMM based mapping model in the sense of the MMSE regression criteria corresponds exactly to the multiple-linear regression model presented in the previous section. Finally, the spectral principal components are sorted following the decreasing order of their explanation variance of the estimated parameter. Then the first p principal components of the spectral parameters as source vectors are according to this order. The parameters of GMM such as mean vectors $\boldsymbol{\mu}_i^{\mathrm{X}}$, $\boldsymbol{\mu}_i^{\mathrm{Y}}$ and the covariance matrices $\boldsymbol{\Sigma}_i^{\mathrm{XX}}$ and $\boldsymbol{\Sigma}_i^{\mathrm{YX}}$ are determined during the GMM training processing. Two different training methods for GMM are discussed in our work. EM is the most used unsupervised training methods for GMM training, which trains the model without any label information of the elements and the data cluster automatically given the initialization parameters by the iterative procedure until the condition of convergence is satisfied [21]. In contrast with the unsupervised training method, a supervised training method is introduced to train the GMM based on the a priori information: the visemes of vowels (see Table 2) which is a speech presentation in the visual domain for the lips or the different Cued Speech five hand positions (see Figure 1) for hand respectively. Thus 3 Gaussian components of GMM corresponding to the three vowel visemes are trained for lips and 5 Gaussian components corresponding to the five hand positions defined in CS are trained for hand.

Table 2. Visemes of French vowels

Visemes	Phonemes of vowels
V1	[a],[i],[e],[ɛ]
V2	[y],[o],[u],[ø]
V3	[ɔ],[œ]

We use the joint source and target vectors defined in equation (7) rather than the source vectors only to train the GMM both in the EM and supervised training methods, which is more robust for small amounts spcifically since the joint density should lead to a more judicious clustering for the regression problem ([20]).

$$\mathbf{z} = [\mathbf{x}, \mathbf{y}] \in R^N \tag{7}$$

Where, N is the dimension of the joint vector $\mathbf{z}$. Once the GMM is trained, the parameters of GMM are fixed.

4.2. Results

The residual variances of the estimated lip and hand parameters obtained by the supervised training GMM decrease significantly compared to the multiple-linear model (see Figure 4). The residual variances decreased with the increment of the dimension of the source vector, i.e. the number of predictors. Finally, the residual variance reaches to around 7% for lip parameters (A, B and S) and 3% for hand coordinates (x, y) respectively by using 16-dimensions source vector.

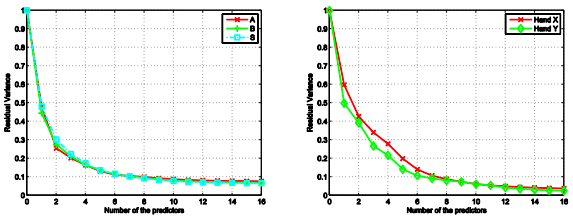

Figure 4: *The average residual variance of the lip parameters and the hand coordinates on the training data in function of the dimension of the source vector. On the left column, the number of the Gaussians m=3 for the lips parameters and on the right column, m=5 for the hand parameters.*

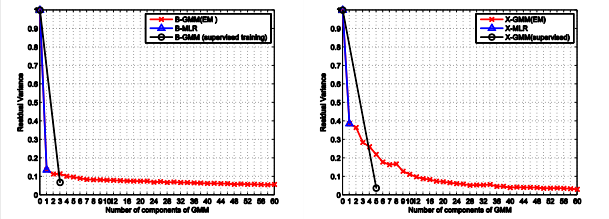

Figure 5: *The average residual variance of the lip parameters B (on left) and the hand coordinates x (on right) on training data in function of the number of Gaussians in GMM. The red line with crosses corresponds to the EM training GMM, the black line with squares corresponds to the supervised training GMM, the blue line with triangles corresponds to the multiple-linear regression model.*

Figure 5 compares the residual variance obtained by the multiple-linear regression model, the supervised training GMM mapping model and the EM trained GMM mapping model. The results show that the residual variance tends asymptotically to a limit value when the number of the Gaussians increases for both lip and hand parameters estimation in the case of EM trained GMM. And the residual variances obtained by the multiple-linear model (denoted by the blue triangle) are completely equal to the ones obtained by the uni-Gaussian GMM model. The supervised training GMM mapping model shows the competitive performance compared to the EM trained GMM model in terms of number of Gaussians. That is to say the supervised training GMM mapping model is more efficient than the one trained by EM method. In addition, the supervised training GMM model also shows good robustness in the evaluation procedure (i.e. test procedure) where it also retains the superior performance (see Table 3).

Table 3. Evaluation results of the lip and hand parameter mapping in terms of the 3 different models.

(%)	MLR	GMM (supervised)	GMM (EM)*
A	17%	9%	7% (38 comp.)
B	12%	9%	9% (40 comp.)
S	14%	8%	8% (40 comp.)

(%)	MLR	GMM (supervised)	GMM (EM) *
X	43%	8%	8% (60 comp.)
Y	31%	4%	5% (54 comp.)

* GMM (EM) shows the minimum residual variance and the number of the Gaussians with which the best results were obtained.

5. Discussion

These successive maps processing of the acoustic spectrum to the lips parameters and the hand position showed that it is much more difficult to estimate the hand position than lips parameters. Several different approaches, from the direct multi-linear method to the sophisticate GMM-based regression method, have been employed to this problem. Actually the source of the difficulty is that there is no relation between the hand position and the spectral parameters unlike the case of the lips as a vocal articulator with corresponding acoustic consequences. More specifically, there are two key points of the meaning of "no relation": (1) there is no structural topological relation between the acoustic space and the hand position space. That is to say, the two closed vowels in the acoustic space may be very far in the hand position space, such as the vowel [e] and [i]. On the contrary, two far vowels in the acoustic space may be corresponding to the same hand position, such as vowel [a] and [o]. It indicates that the two spaces have totally different topology structure since the hand position is determined by the rules of CS but not the acoustic parameters. This is the real reason why a large residual variance was obtained with the multi-linear approach; (2) there is no relation of the variance within group between the acoustic space and the hand position space. That is to say, the tiny variation of the sound will not consequently change the hand position of the speaker. Indeed the hand position around the center within group is random from person to person. Thus it is impossible to establish a global linear relation by the multi-linear model or even a local linear relation by the GMM to predict the movement of the hand around the center within group from the acoustic spectrum.

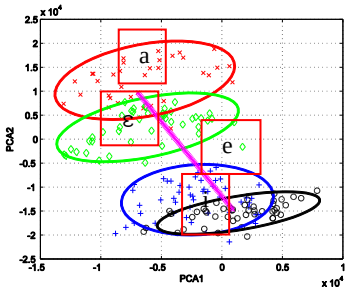

Figure 6. *The linear interpolation in the acoustic space between vowels [a] and [i].*

In order to have a further understanding and comparison of the different mapping approaches, a continuous transition has been achieved by a linear interpolation in the acoustic spectral parameter (i.e. MFCC+LSP) between the vowels [a] and [i]. The 16-dimension acoustic spectral parameters were projected onto their first two PCA components to verify the continuous linear transition in the acoustic space (see Figure 6). In fact there are many ways to go to vowel [i] from vowel [a] in the acoustic space in function of the choice on different starts and ends, but here only one of them is presented as an example to show the corresponding transition obtained by the different mapping models. The corresponding transitions of the estimated lip parameters and hand positions are shown in Figure 7 and Figure 8. For the multi-linear method, the figures present a reasonable linear relation between the linear interpolation spectral parameters and the estimated hand position or lip parameter. For the GMM based mapping method (in the MMSE regression criterion, with 5 components corresponding to the five hand position in CS for hand position estimation and 10 components corresponding to the ten vowels for lip parameter estimation), the four stable phases during the transition both for the hand position and lips are corresponding to the passing vowels ([a],[ɛ],[e],[i]) during the spectral parameters changing linearly from vowel [a] to [i] in the acoustic space. With the four stable phases, the GMM-based mapping method shows classification-like property which helps the model to decrease significantly the residual variance in comparing with the multi-linear model. However, unlike the GMM-based classification method which cannot project the variance of the source data at all, the GMM-based mapping method can still reflect the linear relation locally in the region of the phase as shown in the figures. Due to the strong linear correlation between the acoustic spectrum and the lips parameter, the transition of lips shown in Figure 7 is different as the case of the hand. The order of the phases change in coherence with the lip parameter B and the multi-linear model performs well passing close to the centers of phases corresponding to the different vowels. The local linear regression of GMM-based mapping method is effective and varies in the right direction, however in the case of the hand position estimation the local regression is weak and even in the wrong direction (such as the local regression on the phase of vowel [a] in Figure 6) due to there is "no relation" between the hand position and the acoustic spectral parameters. With the effective local regression, the GMM-based mapping method can improve the estimation performance in comparison with the multi-linear model.

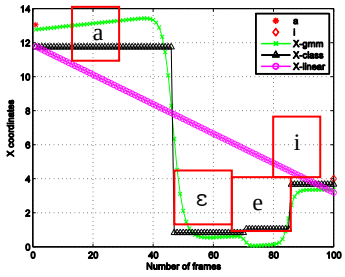

Figure 7: *The dynamic transition of X coordinates of hand position by interpolation between the vowel [a] and [i]. Results with the multi-linear mapping (X-linear), the GMM mapping (X-gmm), and the gaussian classifier applied to the spectral space (X-class).*

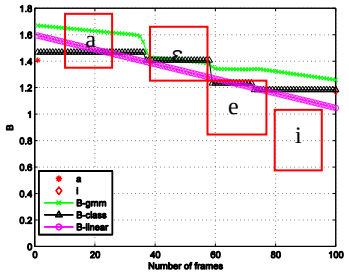

Figure 8: *The transition of lip parameter B by interpolation between the vowel [a] and [i]. Results with the multi-linear mapping (B-linear), the GMM mapping (B-gmm), and the gaussian classifier applied to the spectral space (B-class).*

In the case of the hand positions estimation, the residual variance of the GMM-based mapping approach decreases

significantly comparing to the linear approach by the class-like property but results in the phase changing rapidly. In the case of lip parameters estimation, the local regression of the GMM-based mapping approach is more effective and the phase changing is more gradual meanwhile the multi-linear approach also performs well due to the strong linear correlation between the acoustic spectral parameters and lips parameters.

With both the properties of the classifier and the regression estimator, the GMM-based mapping method decreases the residual variance of the estimated value significantly comparing with the multi-linear model. These properties are well presented in the case of the lip parameters estimation in which the local regression and classification methods perform well. However, when the relation is weak or even no relation between the source and target data, the local regression will degenerate or even meaningless such as the case of the hand position estimation. At this time, the GMM-based mapping method more tends to the classification method, thus the residual variance may be no longer appropriate for evaluating the model since the errors probably cannot be understood at all even if the model has a small residual variance. The cognitive effect of the human being in a perception task may be an alternative evaluation criterion. But note that in the GMM-based mapping method, there is one point essentially different with the classification method that is the contributions of all the components are always considered and weighted to produce the final results. From this aspect, GMM-based mapping method will effect better than the binary classification method in the mapping problem.

6. Conclusion

This paper discusses the relations between the speech spectral space and the visual space of speech and Cued Speech. This program started with the case of oral French vowels. The multiple-linear regression model as a simple case of the GMM modeling has been first used to convert the spectral parameters towards the lip parameters as well as the hand parameters of the Cued Speech. The results show that the best predictors are 16 principal components derived from the 16 LSP and 16 MFCC coefficients. The linear approach showed its limit in the case of the manual setting hand component of the Cued Speech. Two types of GMM based model have been introduced to solve the mapping problem. Since the GMM based model explore the regression relationship between the source and target vectors based on the Gaussians locally and precisely rather than the rough regression based on the global set as in the multiple-linear model, the results obtained by the GMM based model were improved significantly with an explanation of 93% for lip and 96% for hand components of the original variance for the best results. In addition, the supervised training GMM shows a high efficiency and good robustness benefiting from the a priori phonetic information in comparison with the EM training GMM which may be affected more by the outliers due to the important dependence on the data itself. For the future, these results have to be evaluated in perception with deaf persons using supplemented visual speech synthesizers.

7. Acknowledgements

The authors wish to thank the speaker to have accepted the constraints of recording and Thomas Hueber for the fruitful discussions on the method. This work is supported by the National Agency of French Research through the TELMA and PLASMODY projects.

8. References

[1] Montgomery, A. A., Jackson, P. L., 1983. Physical characteristics of the lips underlying vowel lipreading performance. Journal of the Acoustical Society of America 73(6),

[2] Nicholls, G., Ling, D., 1982. Cued Speech and the reception of spoken language. Journal of Speech and Hearing Research 25, 262-269.

[3] Auer, E.T. Bernstein, L.E., 2007. Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment. Journal of Speech, Language, and Hearing Research, Vol.50, pp. 1157-1165.

[4] Bernstein, L.E., Auer, E.T., Jr, & Jiang, J. (2010). "Lipreading, the lexicon, and Cued Speech", in C. la Sasso, J. Leybaert, K. Crain (Eds.), Cued Speech for the Natural Acquisition of English, Reading, and Academic Achievement. Oxford University Press.

[5] Strelnikov, K., Rouger, J., Lagleyre, S., Fraysse, B., Deguine, O. & Barone, P. 2009. Improvement in speech-reading ability by auditory training: Evidence from gender differences in normally hearing, deaf and cochlear implanted subjects. Neuropsychologia, 47, 972-979.

[6] Cornett, R. O. (1967). "Cued Speech," American Annals of the Deaf ,112, 3-13, 1967.

[7] Attina, V., Beautemps, D., Cathiard, M. A. & Odisio, M. (2004). "A pilot study of temporal organization in Cued Speech production of French syllables: rules for Cued Speech synthesizer," Speech Communication, vol. 44, pp. 197-214.

[8] Uchanski, R.M., Delhorne, L.A., Dix, A.K., Braida, L.D., Reed, C.M., Durlach, N.I., 1994. Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech. Journal of Rehabilitation Research and Development 31(1), 20-41.

[9] Leybaert, J., 2000. Phonology acquired through the eyes and spelling in deaf children. Journal of Experimental Child Psychology 75, 291-318.

[10] Tokuda, K., T. Yoshimura, et al. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey.

[11] Zen, H., K. Tokuda, et al. (2004). An introduction of trajectory model into HMM-based speech synthesis. ISCA Speech Synthesis Workshop, Pittsburgh, PE.

[12] Toda, T., A. W. Black, et al. (2004). Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. International Speech Synthesis Workshop, Pittsburgh, PA.

[13] Uto, Y., Y. Nankaku, et al. (2006). Voice conversion based on mixtures of factor analyzers. InterSpeech, Pittsburgh, PE.

[14] Duchnovski, P., D. S. Lum, J. C. Krause, M. G. Sexton, M. S. Bratakos and L. D. Braida (2000). "Development of speechreading supplements based on automatic speech recognition." IEEE Transactions on Biomedical Engineering 47(4): 487-496.

[15] Gibert, G., Bailly, G., Beautemps, D., Elisei, F., & Brun, R. (2005). "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using Cued Speech," J. Acoust. Soc. Am., vol. 118(2), pp. 1144–1153.

[16] Beautemps, D., Girin, L., Aboutabit, N., Bailly, G., Besacier, L., Breton, G., Burger, T., Caplier, A., Cathiard, M.A., Chêne, D., Clarke, J., Elisei, F., Govokhina, O., Le, V.B., Marthouret, M., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Sérignat, J.F., Tribout, M., Vidal, S. (2007), "TELMA: Telephony for the Hearing-Impaired People, From Models to User Tests," In Proceedings of. ASSISTH 2007, pp. 201–208

[17] Berthommier F. (2003). Audiovisual Speech Enhancement Based on the Association between Speech Enveloppe and Video Features. In Proceedings of Eurospeech'2003.

[18] Lallouache, M.T. (1991). "Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres," Ph.D. Thesis, Institut National Polytechnique de Grenoble, 1991.

[19] Kain, A. (2001). High-resolution voice transformation (PhD, OGI School of Science & Engineering, Oregon Health & Science University)..

[20] Hueber, T., Benaroya, E.L, Denby, B., Chollet, G. (2011). "Statistical Mapping between Articulatory and Acoustic Data for an Ultrasound-based Silent Speech Interface", Proceedings of Interspeech, pp. 593-596, Firenze, Italia.

[21] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. New York: Springer.