HMM-based Automatic Visual Speech Segmentation Using Facial Data

Utpala Musti, Asterios Toutios, Slim Ouni, Vincent Colotte,
Brigitte Wrobel-Dautcourt, Marie-Odile Berger

Université Nancy 2 - LORIA, UMR 7503, BP 239, 54506 Vandœuvre-lès-Nancy, France
{musti,toutiosa,slim,colotte,wrobel,berger}@loria.fr

Abstract

We describe automatic visual speech segmentation using facial data captured by a stereo-vision technique. The segmentation is performed using an HMM-based forced alignment mechanism widely used in automatic speech recognition. The idea is based on the assumption that using visual speech data alone for the training might capture the uniqueness in the facial component of speech articulation, asynchrony (time lags) in visual and acoustic speech segments and significant coarticulation effects. This should provide valuable information that helps to show the extent to which a phoneme may affect surrounding phonemes visually. This should provide information valuable in labeling the visual speech segments based on dominant coarticulatory contexts.

Index Terms: facial speech, speech segmentation, forced alignment, coarticulation.

1. Introduction

Speech animation is being explored due to the advantages of audio-visual speech over acoustic speech. Such advantages include improved intelligibility in noisy environments [1], various applications like virtual teaching assistants, language training application for the hearing impaired [2]. The resulting speech animation would be efficient only if it preserves all the information present in natural speech. The most challenging aspect of speech animation is to take coarticulation into account.

To preserve much of the information present in natural visual speech, data based approaches, which include concatenative approaches might be advantageous. Concatenative speech synthesis calls for a well segmented and labeled speech corpus which can be used at the time of synthesis. The segmentation of the corpus for audio-visual speech synthesis should include the information about the asynchrony between visual and acoustic speech segments. In addition, labeling that provides useful coarticulation information on the preceding and succeeding phonemes is beneficial for visual synthesis.

One of the earlier approaches to segment visual data was to assume that visual speech segments are anchored to acoustic boundaries and to use Automatic Speech Recognizers to segment the acoustic speech data [3]. Another approach applied a Hidden Markov Model (HMM) based phasing model which could iteratively learn the time lags between acoustic and visual units, based on the assumption that a unique time lag is associated with each context-dependent HMM [4].

In this paper, we propose an approach to study how well an HMM-based method will capture the uniqueness of visual component of speech. The segmentation is performed using an HMM-based forced alignment mechanism widely used in Automatic Speech Recognition (ASR). The difference between the approach presented in [4] and this one is that, the information regarding the acoustic boundaries was used only to decide when to stop the HMM training and to evaluate the segmentation results. The training is the one typically used in ASR. We train our phoneme HMMs using only visual speech related features. The assumption is that the training would capture the uniqueness in the visual component of speech. In the following section we describe our data and the method used to acquire the visual speech data. Then we describe our alignment method and the 4 experiments used for the evaluation.

2. Data acquisition and representation

We mainly used facial data that is relevant to labial coarticulation, in a broader sense to facial deformations during speech. The data was acquired using a low cost stereo-vision based system [5]. The system uses two fast monochrome cameras, a PC and painted markers that do not change speech articulation. It provides a sufficiently fast acquisition rate to enable an efficient temporal tracking of 3D points. Stereo-vision image pairs of the face painted with 252 markers were captured at a rate of 188.27 Hz simultaneously with acoustic speech. The painted markers were then detected in the image pairs and the corresponding 3D points were reconstructed using a classical stereo-vision algorithm.

The speech corpus has a total duration of about 28 minutes having 319 sentences in French covering 35 phonemes (21 consonants and 14 vowels). This data was sub-sampled to 100 Hz, for easier labeling and alignment with speech-derived acoustic parameters. The sub-sampled data was filtered using a low-pass filter with a cutoff frequency of 25 Hz. We found that such processing removes additive noise from the visual trajectories without suppressing important position information.

We performed our alignment experiments with three sets of features. The first set of feature vectors that were extracted include lip protrusion, lip opening, lip spread and jaw opening (see Fig. 1) [6]. Jaw opening is calculated as the distance between the center of the chin and a fixed point on the head.

The second set of features consists of PCA based parameters. PCA was applied to the markers on the lower face (jaw, cheeks and chin) as presented in Fig. 2. A feature vector of size 7 is obtained by applying PCA transformation based on the first 7 Principal Components which account for about 92% of the data variance. The trajectories of all the markers on the face are not relevant to speech. Markers on the upper part of the face either don't move, or their movements are not relevant to speech. The markers on the lower part of the face are tightly connected to speech gestures.

The third set of feature vectors is the combination of the articulatory and PCA based feature vectors of each facial speech sample, resulting in a heterogeneous feature vector of size 11 consisting of the 4 articulatory features and the 7 PCA based

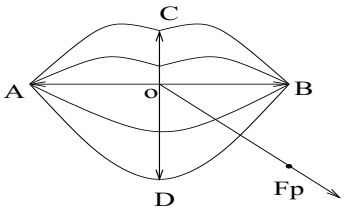

Figure 1: *Calculation of labial features is done using the 4 points on the face: A, B, C and D. Lip opening and lip spread are given by the distances $\|\vec{CD}\|$ and $\|\vec{AB}\|$. Lip protrusion is given by the displacement of O, the center of gravity of the four points (A, B, C, D) along the normal vector ($\vec{OFp}$) to the plane formed by vectors $\vec{AB}$ and $\vec{CD}$*

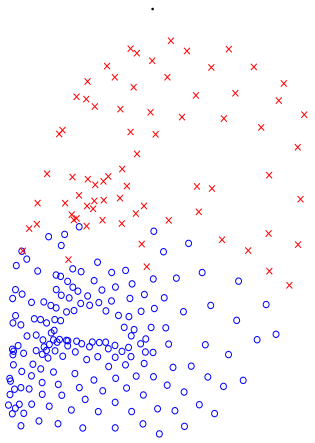

Figure 2: *PCA is applied on 178 (plotted as blue circles) out of 252 painted markers.*

features.

3. Automatic segmentation of visual speech

We trained phoneme HMMs using a procedure similar to the one typically used in ASR [7]. We used the three sets of visual feature vectors as described in the previous section. The set of labels include the set of phonemes covered in the corpus and *sil* (silence).

In the first step, monophone HMMs corresponding to each label were trained. Each HMM was a 3-state left-to-right no-skip model. The output distribution of each state was a single Gaussian with a diagonal covariance matrix. The observation vectors input to the HMM training consisted of static and dynamic parameters, i.e. the three types feature vectors described in the previous section and their delta and delta-delta coefficients. The HMM parameter estimation was based on the ML (Maximum-Likelihood) criterion estimated using Baum-Welch recursion algorithm. The learned monophone HMMs were used to perform a forced alignment of the same training corpus. The set of monophone HMMs, among the three sets trained on the three different feature vectors, which gave the best result based on the total recognition error criterion (explained later in this section) was chosen for the second step in which the segmentation is improved further. The second training step involved creation of context dependant triphone models using the trained monophone HMMs and finally the creation of tied-state triphones using decision tree clustering. The triphone

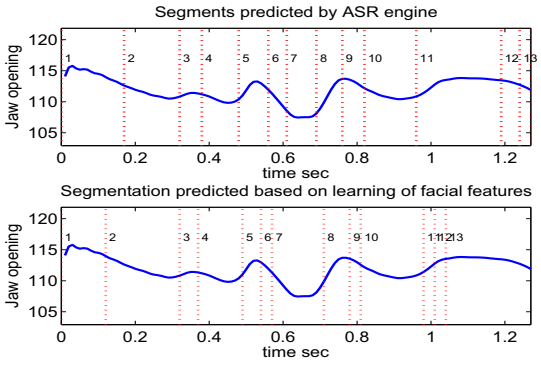

Figure 3: *Segments predicted by the ASR engine and our method which is based on learning from the facial features for the phoneme sequence "sil s e z a d v ɛ ʁ s ɛ ʁ sil".*

models were created by cloning the trained monophone HMMs for different triphones and re-estimating the HMM parameters for those triphones contexts which can be reliably estimated using the corpus. Then using decision tree clustering tied state triphones were created. The contexts considered for clustering are based on the hierarchical cluster trees of phonemes mentioned in [8]. The complete speech corpus has been used for the estimation of HMM parameters. These trained HMMs were then used to perform forced alignment of the data.

It has been shown that visual speech segments are correlated to the corresponding acoustic speech that has to be produced [9, 10]. Thus there has to be an overlap between the actual acoustic segment and visual speech segments. In fact, the speech sound is the consequence of the vocal tract deformation and thus the face. The visual and acoustic speech segments might have asynchrony in their onset and end time as the vocal tract has to anticipate the following sound by adjusting the different articulators.

Based on the above reasoning of asynchrony and overlap of the visual and acoustic speech, the following criterion has been derived for evaluating the segmentation results. The recognition of any label would be considered correct if there is an overlap between the predicted visual segment and the actual acoustic segment, the overlap being however small. An ASR engine was used to provide the phoneme labels and acoustic boundaries of the whole corpus. We consider the acoustic boundaries given by the ASR engine are the correct acoustic boundaries. After each iteration of HMM parameter re-estimation, the training data is segmented using the updated HMMs and the total recognition error of the segmentation are calculated. Training is halted when there is no further improvement in this value. The recognition error of each labeled visual segment in the corpus at this stage has been used for the evaluation and analysis of the alignment results (see Fig. 3 for an example).

4. Forced alignment results

In this section, we present the quantitative results based on the recognition error mentioned in the previous section. We classify phonemes based on their visibility as shown in Table 1. We consider /ɥ/, /w/, /ʃ/ and /ʒ/ as bilabial based on their secondary place of articulation. In fact, their primary place of articulation is not relevant to our study (not visible) as it is the case

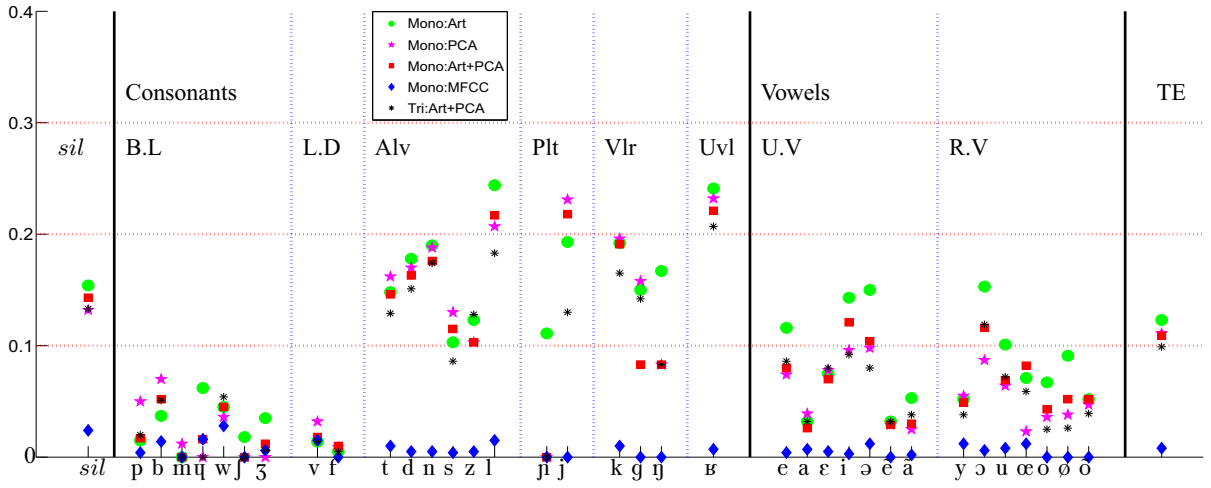

Figure 4: *Recognition errors in the alignments: Art, PCA are the articulatory and principal component analysis based feature vectors; MFCC are the acoustic parameters (Mel-frequency cepstral coefficients); Mono and Tri are the monophone and triphone HMMs. TE is the total recognition error.*

Vis.	Abbr.	Class	Members of the class
1	B.L	bilabial	p, b, m, ɰ^{o}, w^{o}, ʃ^{o}, ʒ^{o}
	L.D	labiodental	v, f
	R.w	rounded vowels	y, ɔ, u, œ, o, ø, õ
2	*sil*	*sil*	*sil*
	Alv	alveolar and dental	t, d, n, s, z, l
	Plt	palatal	ɲ, j
	vlr	velar	k, g, ŋ
	Uvl	Uvular	ʁ
	U.V	unrounded vowels	e, a, ɛ, i, ə, ẽ, ã

Table 1: *Classification of phonemes based on their visibility. Phonemes classified as 1 are visible and 2 are invisible. Phonemes followed by ○ are classified based on their secondary place of articulation.*

for the secondary place of articulation. A three-way classification of phonemes has been proposed based on their importance of lip-reading as invisible, protected and normal [11]. Invisible phonemes are those which are not associated to a specific shape of the mouth during their articulation and thus their articulation is not visible from outside the mouth. Protected phonemes are those whose characteristic coarticulation effects must be preserved in lip motion synthesis, like the phoneme /p/ (where lips should be completely close).

We performed 4 alignment experiments. These include 3 experiments based on training monophone HMMs using the 3 types of feature vectors mentioned in Section 2. Based on the alignment results with the 3 sets of monophone HMMs, the feature vector performing the best among the three was selected for training the context dependant triphone models for further improvement of alignment. The results are presented in Figure 4.

The PCA based feature vectors perform better than articulatory feature vectors in terms of the total recognition error. The heterogeneous feature vector, consisting of both PCA based features and articulatory features, performs better than each taken alone. PCA based features quantitatively account for the overall shape or deformation during the speech production and the articulatory parameters increase the discrimination by quantifying the typical articulatory characteristics like complete closure of mouth for /p/. This performance is further improved by triphone HMMs. As one can expect the recognition errors are low for protected phonemes or the phonemes which involve labial region for their coarticulation. The recognition errors are relatively higher for other consonant classes.

To verify that substantial training can be achieved by our small corpus (28 minutes of audio-visual speech), monophone HMMs were trained using the acoustic speech of our corpus. The acoustic features extracted from the speech were the MFCC (Mel-frequency cepstral coefficient) features. The trained HMMs were used for the forced alignment of the same speech data that was used for training. The resulting acoustic segments were compared with the segments predicted by the ASR engine. The total recognition error used to quantify the visual segmentation results was determined in this case. A total recognition error of $< 1\%$ was observed.

The following analysis has been done considering only the correctly recognized visual segments. Let As and Vs be the starts of the acoustic and visual segments of a label, Ae and Ve be the ends of the acoustic and visual segments of the label. Let Ds be the start difference and De be the end difference, calculated as follows:

$$Ds = (As - Vs),$$
$$De = (Ae - Ve)$$

The mean and variance of Ds and De are calculated for each of the labels covered by the corpus (see Fig. 5 and Fig. 6). A positive expectation of the start difference, $(E(Ds) > 0)$ means visual start leads over the acoustic start which suggests a visual influence of the speech coarticulation on the left contexts. Based on the segmentation results this is the case for bilabials, labiodeltals and rounded vowels. Similarly $(E(De) < 0)$ means acoustic end leads over visual end with a visual influence of the speech coarticulation on the right context. The segmentation results obtained show that /ɰ/, /w/, /ʃ/ and /ʒ/ fall in this

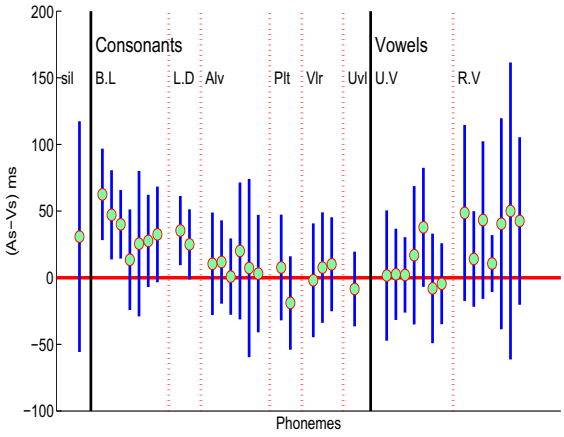

Figure 5: *Mean difference in the starts of acoustic and visual speech segments*

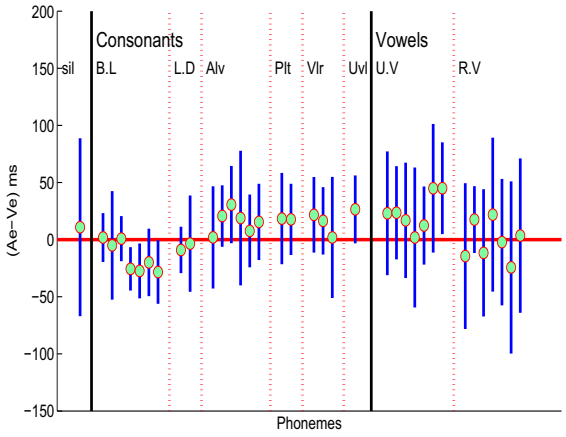

Figure 6: *Mean difference in ends of acoustic and visual speech segments*

category.

5. Discussion

A concept which is important in this context is the "deformability" of phonemes defined as the extent to which a phoneme coarticulation can be modified by surrounding phonemes [12]. The coarticulation of phonemes captured by the facial speech data in the presence of deformations can be labeled based on the characteristic of the dominant deformation based on surrounding phoneme contexts (like an influential /u/ deformation on /p/ in 'pu'). Such a labeled visual speech corpus might be helpful for better concatenative results. The influence of neighboring phonemes can be assumed to be captured by the triphone models. The automatic characterization or labeling of coarticulation effects related to deformation beyond just the neighboring phoneme boundaries like in the word 'strewn' where /s/ and /t/ have the anticipatory lip rounding is challenging [13]. Characterizing the phoneme segments based on these deformations might be useful in synthesis of a better speech animation. The usage of segments like the /s/ of 'strewn' for concatenative synthesis at contexts without a rounding which is contrary to its characteristic deformation in the actual corpus would be undesirable. The segmentation of visual speech using facial speech features might pave way towards automatic labeling of visual speech segments based on their deformability. This aspect of the segmentation algorithm might be explored further to arrive at systematic qualifying conclusions. For improving the segmentation results mainly for invisible phonemes the addition of tongue related data (using Electromagnetic Articulography) can also be explored.

6. Acknowledgement

This work was supported by the French National Research Agency (ANR - ViSAC - Project N. ANR-08-JCJC-0080-01).

7. References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[2] D. Massaro, "Embodied agents in language learning for children with language challenges," *Computers Helping People with Special Needs*, pp. 809–816, 2006.

[3] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, "Accurate automatic visible speech synthesis of arbitrary 3D models based on concatenation of diviseme motion capture data," *Computer animation and virtual worlds*, vol. 15, no. 5, pp. 485–500, 2004.

[4] O. Govokhina, G. Bailly, and G. Breton, "Learning optimal audiovisual phasing for a HMM-based control model for facial animation," in *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW6-2007)*, Bonn, Germany, 2007.

[5] B. Wrobel-Dautcourt, M. O. Berger, B. Potard, Y. Laprie, and S. Ouni, "A low-cost stereovision based system for acquisition of visible articulatory data," *In AVSP-2005*, pp. 145–150, 2005.

[6] V. Robert, B. Wrobel-Dautcourt, Y. Laprie, and A. Bonneau, "Inter speaker variability of labial coarticulation with the view of developing a formal coarticulation model for french," *In AVSP*, pp. 65–70, 2005.

[7] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Engineering Department, 2005.

[8] M. Odisio, G. Bailly, and F. Elisei, "Tracking talking faces with shape and appearance models," *Speech Communication*, vol. 44, no. 1-4, pp. 63–82, 2004.

[9] J. P. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS '99)*, 1999.

[10] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[11] S. Minnis and A. Breen, "Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis," in *Sixth International Conference on Spoken Language Processing*, vol. 2, Beijing, China, 2000, pp. 759–762.

[12] C. Pelachaud, N. I. Badler, and M. Steedman, "Linguistic issues in facial animation," in *Computer Animation*, vol. 91, 1991, pp. 15–30.

[13] R. D. Kent and F. D. Minifie, "Coarticulation in recent speech production models," *Journal of Phonetics*, vol. 5, no. 2, pp. 115–133, 1977.