A study on auditory feature spaces for speech-driven lip animation

Guylaine Le-Jan, Yannick Benezeth, Guillaume Gravier, Frédéric Bimbot

A study on auditory feature spaces for speech-driven lip animation

Guylaine Le-Jan, Yannick Benezeth, Guillaume Gravier, Frédéric Bimbot

IRISA/CNRS & INRIA
Campus de Beaulieu, 35042 Rennes, France
guylaine.lejan@inria.fr

Abstract

We present in this paper a study on auditory feature spaces for speech-driven face animation. The goal is to provide solid analytic ground to underscore the description capability of some well-known features with relation to lipsync. A set of various audio features describing the temporal and spectral shape of speech signal has been computed on annotated audio extracts. The dimension of the input feature space has been reduced with PCA and the contribution of each input feature is investigated to determine the more descriptive. The resulting feature space is quantitatively and qualitatively analyzed for the description of acoustic units (phonemes, visemes, *etc*.) and we demonstrate that the use of some low-level features in addition to *MFCC* increases the relevance of the feature space. Finally, we evaluate the stability of these features *w.r.t.* the gender of the speaker.

Index Terms: features selection, description of acoustic units, lipsync.

1. Introduction

The visual component of speech provides valuable information that undoubtedly increases the intelligibility of speech. Visible speech can compensate for a substantial loss in the speech signal and can also compensate a degraded audio component [1]. The intelligibility of natural speech can also be increased with synthetic faces. It is also important to note that a synthetic face with a realistic lip animation increases the human-computer interaction. In order to obtain natural looking and realistic animation of virtual characters, the lip movements shall be consistent with the corresponding speech. It is possible to obtain such an animation manually by adjusting, frame after frame, the control parameters of the synthetic face lips. The results are then very realistic, but this method requires a substantial workload and does not allow real time avatar animations. Usually, methods to automatically generate lip movements are based on speech recognition approaches. In its basic form, the synchronization of lip movements with the speech signal implies the detection of the phoneme sequence (*e.g.*, [2]). Phonemes are then mapped to a set of visemes. A viseme is a lip configuration of the animated character. The recognition system can also be trained to directly detect visemes (*e.g.*, [3]). In both cases, the mouth movements are monitored with a discrete set of positions: the visemes.

Because of the large number of applications that require speech-lip synchronization, it is now a well-studied problem. In human-computer interactions, virtual characters with realistic lip animation allow to interact with users in a friendly and natural way. Video games, eLearning, instant messaging or 3D animation productions (movies, commercials, *etc*.) need also to generate lip animations. All these lipsync applications have different constraints in terms of computational resources, real-time or animation quality. A virtual character in mobile device interfaces can obviously tolerate an approximate lip animation but needs real-time animations with high computational contraints. The opposite observation can be drawn concerning the 3D animation productions.

As explained previously, a lipsync system is often based on the recognition of a speech acoustic unit, *i.e.*, determining a sequence of phonemes or visemes. In that case, the recognition system can be broadly divided into two main components: the feature extraction step and the classification step. HMM [4], neural networks [5], Bayesian estimation, mixture models of Gaussian or vector quantization can be used to map acoustic vectors to the corresponding sequence of acoustic units. This classification step can employ a variety of feature spaces. This representation shall characterize the information carried by the acoustic signal in a convenient form. The well-known mel frequency cepstral coefficients (*e.g.*, [4], [5]), the LPC, the Rasta-PLP, or even low-level features such as the zero crossing rate or the energy [6] can be used. In this paper, we are interested in measuring the contribution of low-level features to lip-position in order to design an optimal feature space for lipsync. The separability of the desired classes (set of visemes or phonemes) in a given feature space is not evaluated based on a recognition rate obtained with a classification but the feature space is directly analyzed.

Feature selection methods have been proposed in the literature (*e.g.*, [7], [8]). However, these methods aim at selecting features for speech recognition applications. As far as we know, no such study is dedicated to lipsync systems. This experimental study will helpfully contribute to the acoustic feature selection during the design process of a speech-driven face animation system. We present in this paper a selection of the more descriptive features with PCA. The quality of the determined feature space is then evaluated with regard to the separability of the given classes. The stability of these features considering the gender of the speaker is also considered.

We first present in this paper the experimental protocol with the description of our speech database and the features analyzed. Then, we present a method to identify a new feature space and we evaluate qualitatively and quantitatively the selected feature space.

2. Experimental protocol

2.1. Speech database

The auditory features validity has been performed on 6 continuous speeches of French native-speakers (3 male and 3 female speakers) extracted from the CPROM corpus [9]. Two phonetic experts have annotated this corpus following a strict protocol. The duration of the corpus is about 11 minutes for male and female speakers. The reason for choosing audio extracts from the CPROM corpus rather than from larger databases relies in the high accuracy of the phonemic annotation.

We use a set of 33 phonemes to describe the French language. These phonemes are listed in Tables 1 and 2 using the *X-SAMPA* notation. We also use a set of visemes composed of 17 classes (16 plus the neutral viseme). This viseme classification has been performed by Benoît et al. [10] and is illustrated in Table 1.

Table 1: Phonemes aggregated into 16 visemes

Visemes	*Phonemes*
V1	a
V2	i
V3	y, u, 2, 9, o, o~, @, H
V4	e, E, U~
V5	O
V6	a~
V7	p,b,m
V8	t,d,n
V9	k,g
V10	f,v
V11	s,z
V12	S,Z
V13	L
V14	R
V15	w
V16	J

We also present in Table 2, nine classes of phonemes aggregated according to the mode of articulation (oral or nasal vowels, voiced or voiceless plosives, voiced or voiceless fricatives, liquids, nasal occlusion and mid-vowels). This classification will be used in the results analysis section.

Table 2: Phonemes aggregated into 9 articulation modes.

Classes	*Phonemes*
oral vowel (OV)	a, O, o, u, 9, 2, @, y, E, e, i
nasal vowel (NV)	a~, o~, U~
voiced plosive (VP)	b, d, g
voiceless plosive (VLP)	p, t, k
voiced fricative (VF)	v, z, Z
voiceless fricative (VLF)	f, s, S
liquid (L)	R, l
nasal occlusion (NO)	m, n
mid-vowel (MV)	j, H, w

2.2. Acoustic speech features

We present in this part the list of features analyzed. A large set of features has been used. They characterize the temporal or the spectral shape of the speech signal, and also the energetic and perceptual features:

- Temporal features:
  - Zero crossing rate (*ZCR*),
  - Envelope shape statistics: centroid, deviation, skewness and kurtosis (ESS_i),
- Spectral features:
  - Fundamental frequency (f_0),
  - The 4 first formants (f_i) and bandwidths (b_i),
  - Ratio of formant frequencies: f_2/f_1, f_3/f_1, f_4/f_1,
  - 39 *MFCC* (13 static coefficients, and their first and second order derivatives),
  - Spectral shape statistics: centroid, deviation, skewness and kurtosis (SSS_i),
  - Spectral slope (*SS*),
  - Spectral variation (*SV*),
  - Spectral flux *(SFX*),
  - Spectral roll-off (*SR*),
  - Spectral decrease (*SD*),
- Energy features:
  - Energy (*en*),
  - Log of energy (*logen*),
- Perceptual features:
  - Perceptual sharpness (*PSH*).

More information about these auditory features can be found in [11]. Because of the large number of auditory features, it was interesting to propose a new feature space with a lower space dimension and select the more relevant one for acoustic unit description.

3. Feature space definition

We present in this part a selection of an optimal feature space to describe acoustic units. For this, a PCA has been performed in order to analyze the description capability of each feature and reduce the multidimensional data set. At the opposite of other data analysis method, such as LDA, PCA disregards the information on the class of each example. This method is more appropriate for our study because the analysis is not dedicated to the specific problem of phoneme recognition. PCA find the projection that maximizes the variance of the projected data and provide principal components defined as linear combination of the original variables. Each extracted factor can be interpreted through the loadings of the initial variables. The first factors are the most representative of the total variance of the original data. In order to reduce the feature space dimension, we retain only the most contributive factors. In the next part of the study, we retain the first twenty factors such as they explain near 70% of the total variance.

Then, to determine the best set of auditory features, the contribution (*CT*) of each feature is calculated on the set of selected principal factors (such as 20 factors). The contribution CT_i of the feature i to the N selected principal factors is defined as follows:

$$CT_i = \frac{\sum_{\alpha=0}^{N} x_{i\alpha}^2}{\sum_{\alpha=0}^{N} \lambda_\alpha}, \quad (1)$$

where $x_{i,\alpha}$ is the coordinate of feature i on the axis α and λ_α the eigenvalue of the principal factors α.

Features with a high *CT* show a significant contribution to the realization of the new feature space and also to the acoustic unit description. The 10 first features, with the best *CT* are, in decreasing order: ESS_4, ESS_3, *SS*, *PSH*, SSS_1, f_4/f_1, f_3/f_1, SSS_2, *logen*, f_2/f_1. On the other hand, the 10 less contributive features, other than *MFCC*, are: *en, SFX,* B_1, f_0, *SD*, SSS_4, B_4, B_3, B_2, ESS_2.

4. Feature space analysis

We have defined in the previous section a feature space composed of the first twenty factors of the PCA. In order to evaluate the relevance of this feature space, we first present in this part a quantitative comparison of this feature space with the original feature space composed of all auditory features listed in section 2.2 and the usual MFCC-based feature space. Then, we present a qualitative analysis about the observations disposition in the feature space. The objective is to detect outliers and potential problems. We also evaluate the stability of this feature space *w.r.t.* the gender of the speaker

4.1. Quantitative analysis

First, we compare the capacity of the contributive factors to aggregate observations into phonemes, articulation modes and visemes. To do that, we compute features on the annotated speech database. We obtain several observations for each acoustic unit and conserve only the median value to characterize it. Consequently, we have a distribution of elements that belong to various classes (phonemes, articulation modes or visemes). We use the compactness value α and the separation value β defined as:

$$\alpha = \frac{1}{k}\sum_{i=0}^{k}\left(\frac{1}{n_i}\sum_{j=0}^{n_i}\left\|x_{i,j} - x_i\right\|^2\right), \quad (2)$$

$$\beta = \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j=1, j\neq i}^{k} e^{-\left\|x_i - x_j\right\|^2}, \quad (3)$$

where k is the number of classes, n_i the number of observations in class i, $x_{i,j}$ the observation j in class i and x_i the centroid of the class i. The compactness α is a generalized definition of the variance of a vector dataset and the separation β is the pair-wise distance among the class centroids. As the objective is to minimize the distances among the data points in individual classes and to maximize the distances between classes, the smaller α is, and the bigger β is, the easier it is to separate acoustic observations into the desired classes.

We present in Table 3 the compactness values of the phoneme, viseme and articulation mode classes in different feature spaces. We present in Table 4 the separations values. We compare the original feature space of dimension 68 (called FS1) composed of all features listed in section 2.2, the usual feature space composed of the 39 MFCC (called FS2), and the feature space composed of the twenty most representative factors (called FS3).

The smaller the compactness is, and the higher the separation is, the easier it is to separate acoustic observations into appropriate classes. Consequently, few observations can be drawn from the Tables 3 and 4. First, it is relevant to note that the visemes are better distributed than the phonemes within each space. This observation has already been demonstrated (e.g., [12]) from the recognition error rate of phonemes and visemes. Here, we validate this observation from the feature space analysis. Secondly, the articulation mode representation obtains good separability and compactness values. This result is intuitive because observations in the same articulation mode class are likely to share the same acoustic characteristics. Nevertheless, the major differences between Table 1 and 2, show that the articulation mode information is not sufficient for lip animation. This information can still be used to constraint the mouth position with regard to the current articulation mode.

Table 3: Compactness values

	Phonemes	*Visemes*	*Articulation modes*
FS1	0.010	0.004	0.003
FS2	0.007	0.003	0.002
FS3	0.005	**0.002**	**0.001**

Table 4: Separation values

	Phonemes	*Visemes*	*Articulation modes*
FS1	0.167	0.210	0.275
FS2	0.312	0.354	0.448
FS3	0.376	**0.422**	**0.498**

Then, another interesting observation is that the use of low-level features in complement to the usual *MFCC* does not directly improve the feature space representation quality. Actually, the separatbility of the analyzed acoustic units in Table 3 and 4 presents better result in the MFCC-based feature space. However, the dimension reduction with PCA maximizes the variance of the projected data and consequently conserves information while the dimension reduction optimizes the feature space representation.

4.2. Qualitative analysis

In order to analyze in depth the disposition of the observations in the feature space, we perform a hierarchical clustering onto the reduced feature space. We aggregate phonemes classes with a bottom-up agglomerative hierarchical clustering [11] with the average linkage strategy. The distance between each phoneme is the Euclidian distance. The goal is twofold: first we want to evaluate if the phoneme aggregation is related to the articulation modes and then graphically detect if some phonemes have an unexpected behavior. We present in Figure 1 the dendrogram illustrating the phoneme clustering in the reduced feature space.

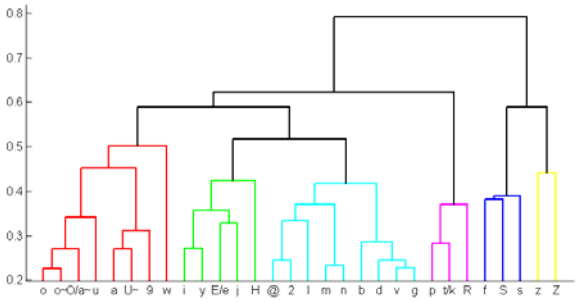

Figure 1: Dendrogram illustrating the clustering of phonemes in the reduced feature space.

Clearly, phoneme aggregation is closely related to the articulation modes. Observations in the same articulation mode class are likely to share the same acoustic characteristics. Phonemes aggregate to form clusters composed of vowels, voiced and voiceless plosives, voiced and voiceless fricatives and nasal occlusions. Nevertheless, the distinction between oral and nasal vowels is not clear. Then, it is possible to observe that liquids are spread into several groups of phonemes and specific phoneme *v* is especially ill represented in the distribution.

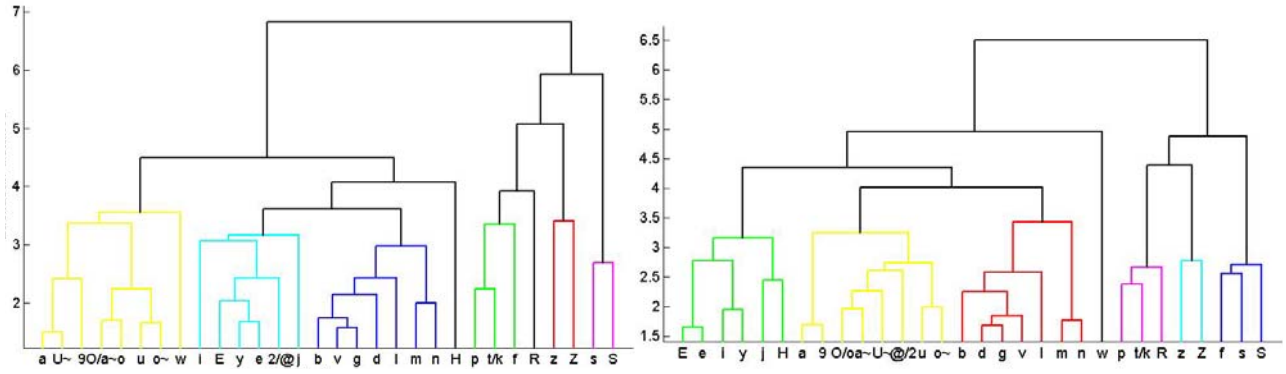

Figure 2: Dendrogram illustrating the clustering of phonemes in the reduced feature space with male speakers (on the left) and female speakers (on the right).

Then we present in Table 5 and in Figure 2, an analysis on the influence on the speaker gender. The Table 5 presents the compactness and the separation values of the phoneme classes on the reduced feature space while the Figure 2 presents the clustering of phonemes also on the reduced feature space.

Table 5: Compactness and separations values illustrating the influence of the speaker gender.

	Males	*Females*
Compactness	**0.004**	0.005
Separation	0.383	**0.404**

First, the compactness and separation values show a slight influence of the speaker gender. While the compactness value is nearly equal, the separation value is higher with female speakers. With a male speaker, mid-vowels are spread into several phoneme classes and phoneme *f* is unexpectedly associated with the voiceless plosives.

This qualitative analysis permits to draw some generalities about the design of speech-driven lip animation systems. First, we again confirm the relevance of the reduced feature space. A feature space composed of the usual *MFCC* with some complementary low-level features, after dimension reduction, characterizes in a coherent form the phoneme groups. Then, we identify some specific phonemes that are inconsistently represented in the feature space. The resulting lip animation would be consequently unrealistic for these occurrences. Nevertheless, the a priori knowledge of these difficulties should influence the design process of lip animation systems. For example, the weight of the phonetic context of these phonemes could be a priori increased to reduce the negative impact on the lip trajectories.

5. Conclusions

We have presented in this paper a study on feature spaces for speech-driven face animation. A set of various audio features describing the temporal and spectral shape of speech signal has been computed on annotated audio extracts. The dimension of the input feature space has been reduced with PCA and the relative contribution of the input features is investigated. The resulting feature space is quantitatively and qualitatively analyzed in a speech-driven animation context. We quantitatively and qualitatively demonstrate that the use of some low-level features in addition to *MFCC* increases the relevance of the feature space when the dimension of this feature space is reduced. We also evaluate the stability of these features *w.r.t.* the gender of the speaker.

6. References

[1] C. Benoît, T. Mohamadi and S. Kandel, "Audio-Visual Intelligibility of French Speech in noise", *Journal of Speech and Hearing Research*, n°37, 1195-1203, 1994.

[2] J. Park and H. Ko, "Real-Time Continuous Phoneme Recognition System Using Class-Dependent Tied-Mixture HMM With HBT Structure for Speech-Driven Lip-Sync", *IEEE Transactions on Multimedia*, Vol. 10(7), pp. 1299–1306, 2008.

[3] S.-W. Foo and L. Dong, "Recognition of Visual Speech Elements Using Hidden Markov Models ", *Advances in Multimedia Information Processing*, Vol. 2532, pp. 153–173, 2002.

[4] J. Park and H. Ko, "Real-Time Continuous Phoneme Recognition System Using Class-Dependent Tied-Mixture HMM With HBT Structure for Speech-Driven Lip-Sync", *in IEEE Transactions on Multimedia*, Vol. 10(7), pp. 1299-1306, 2008.

[5] G. Zorić and I.-S. Pandžić, "Real-time language independent lip synchronization method using a genetic algorithm" *Signal processing*, Vol. 86(12), pp. 3644-3656, 2006.

[6] M. Malcangi and R. de Tintis, "Audio Based Real-Time Speech Animation of Embodied Conversational Agents ", *Gesture-Based Communication in Human-Computer Interaction,* Vol. 2915, pp. 429-440, 2004.

[7] E.-L. Bocchieri and J.-G. Wilpon, "Discriminative feature selection for speech recognition", *Computer Speech & Language* Volume 7, Issue 3, July 1993, Pages 229-246

[8] G. Saon and M. Padmanabhan. "Minimum Bayes error feature selection for continuous speech recognition", In *Advances in Neural Information Processing Systems 13*, pages 800–806, 2001.

[9] M. Avanzi, A.C. Simon, J.-P. Goldman and A. Auchlin, 2010. "C-PROM. An annotated corpus for french prominence studies", *Proceedings of Prosodic Prominence: Perceptual and Automatic Identification, Speech Prosody,* 2010.

[10] C. Benoit, T. Lallouache, T. Mohamadi and C. Abry, "A set of French visemes for visual speech synthesis", *Les cahiers de l'ICP*, Vol. 3, pp. 113-129, 1994.

[11] G. Peeters, *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*, 2004

[12] E. Bozkurt., C.-E. Erdem, E. Erzin, T. Erdem, M. Ozkan, "Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation", *3DTV Conference*, 2007.

[13] A. Jain, M. Murty and P. Flynn, "Data clustering: a review", *ACM computing surveys*, vol. 31(3), pp. 264-323, 1999.