A PROBABILISTIC MODEL OF DOUBLE-VOWEL SEGREGATION

Laurent Varin and Frédéric Berthommier

Institut de la Communication Parlée / INPG
Grenoble, FRANCE
{varin,bertho}@icp.grenet.fr

ABSTRACT

The decomposition principle was first proposed by Varga and Moore [1] and applied to Automatic Speech Recognition (**ASR**) in noise. We show a new adaptation of this principle to model the schema-based streaming process which was inferred after psychoacoustical studies [2]. We address here the classical problem of double vowel segregation. The signal decomposition is allowed by an internal and statistical model of vowel spectra. We apply this decomposition model able to reconstruct the spectra of superimposed signals after identification of only the dominant or of both members of the pair. Three stages are invoked. The first one is a module performing identification when the input is a mixture of interfering signals. Prior identification of the dominant spectra prevents combinatorial reconstruction. The second step is an evaluation of the mixture coefficient also based on an internal representation of spectra. Finally, the reconstruction of spectra is probabilistic, by the way of likelihood maximisation. It uses labels and mixture coefficient. This is tested on a large database of synthetic vowels.

1. INTRODUCTION

In the context of Computational Auditory Scene Analysis (**CASA**), grounded by research achieved by psychologists [2] and auditory modellers, the "cocktail party" problem has been *simplified* to get more insight. In this way, a major paradigm has been first proposed by Scheffers [3]. This reduced task consists in identifying *both members* of a pair of stationary and harmonic vowels. This paradigm mainly concerns the simultaneous organisation of the auditory scene (e.g., signals emitted at the same time).

A common cut used in the CASA domain concerns the differentiation between *primitive* and *schema-based* levels. Primitive segregation processes use cues such as harmonicity and spatial localisation. Specialised representations of signals are built in order to segregate and group phonetic features such as vowel formants when underlying harmonicity or common Interaural Time Differences exists. On the contrary, the current schema-based segregation process is directly applied on the spectral (auditory like) representation where phonetic features are mixed. It uses *memorised* information and works at the phonetic level. This is still compatible with the primitive segregation approach, based on AM map representation [4]. Both techniques are able to evaluate contribution of each source within channels at a given time: these are "shared-channels" segregation techniques. One project is to *couple* them in order to enhance segregation.

In the context of robust ASR, methods have been proposed to *decompose* signal and noise contribution [5]. This had been adapted to HMMs in order to track two sources at the same time [1]. To *combine* CASA and ASR approaches, we apply a general principle elaborated to process information to the double vowel perception task. A previous model able to perform double identification [6] (this better corresponds to the psychoacoustical task) is completed by a recovering process of spectra, so that decomposition of vowel mixtures is now *effective*.

2. MODELLING PRINCIPLE

2.1 Identification and reconstruction

Having a mixture of two vowels, the goal is to label both, and to achieve an optimised reconstruction of the original spectra. Knowing the statistical distribution of all vowels allows us to assign a probability to each spectrum we have in hand during the segregation process. When input is a sum of two signals with spectra s_1 and s_2, the two main stages are:

- An *identification* level, giving the classes C_1 and C_2 of members of input mixtures, or optionally, just the class of the dominant member.
- A *reconstruction* level, getting one or two labels and producing two spectra x_1 and x_2, which maximise probability $P(x_1|C_1)*P(x_2|C_2)$, according to the variance/covariance matrix of the data set.

This corresponds to the likelihood maximisation constraint. This maximisation process allows to find the most probable pair of spectra, knowing the sum, which is the input signal, and supposing the two classes of signals are represented in the reference database. The underlying condition is the *independence* of occurrence and production of the two signals. This is physically assumed when phonemes emitted by different speakers at the same time.

2.2 Evaluation of the mixture coefficient

To apply the maximum likelihood constraint, an other parameter is needed: the relative intensity (e.g., the mixture coefficient). This can be: **(1)** fixed arbitrarily (or given) **(2)** externally given **(3)** evaluated at the primitive level (for. ex., by the way of the "old plus new" heuristic) **(4)** estimated at the classification level. Moreover,

estimation can be re-iterated in a second pass after reconstruction.

We have explicitly defined a stage allowing evaluation of the mixture coefficient. Thus, the system has three steps (Figure 1).

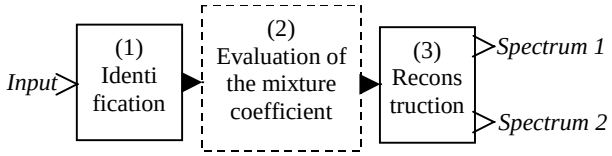

Figure 1: The processing steps

2.3 The dominance effect

A double-vowel spectrum presented to the identification system is associated with the label of the *dominant* vowel. This dominance effect is due to both the relative level and the distribution of spectral features (formants) across the database. When a vowel has a specific formant, this is never masked, and this will be more salient relatively to another one at the classification level. Notice that a restricted definition of dominance has been stated for primitive processes, depending on relative level in each channel [7].

Correct identification of the dominant vowel allows to disrupt the *combinatorial* processing of pairs. This differentiates our model from Varga and Moore's [1] proposal. In this work, we have developed two heuristics:

- **H1**: Double-identification is performed by step (1), so that only a pair is processed during further steps.
- **H2**: Only the dominant vowel is identified first, all possible second members are processed in step (2) and (3). Finally the best pair is chosen, so that a *late identification* of the second vowel is performed.

We briefly present the double identification model, which is an adaptation of the Gaussian classifier able to tackle with the mixtures identification. The late identification strategy (e.g., H2) will be used during further simulations of the complete system.

3. DOUBLE-IDENTIFICATION

The identification is performed with a classical Gaussian classifier. Practically, 30 dimensions represent a spectrum between 100 and 5000Hz in Mel scale. The "training" stage consists to compute the variance/covariance matrix (having dim. (30,30)) of isolated vowels.

The principle of mixture recognition is first to recognise the dominant vowel and to do a copy of the input spectrum. We get the second label by removing *iteratively* and *linearly* the mean spectrum of the dominant vowel class is this copy. This enables the second object to be enough *unmasked* to be recognised. Conversely, the second vowel can be suppressed in the first copy, and this leads in a few cases to get another label. Figure 2 shows the identification of the pair /a/+/i/. The recognition rate of the pairs in a set of synthetic

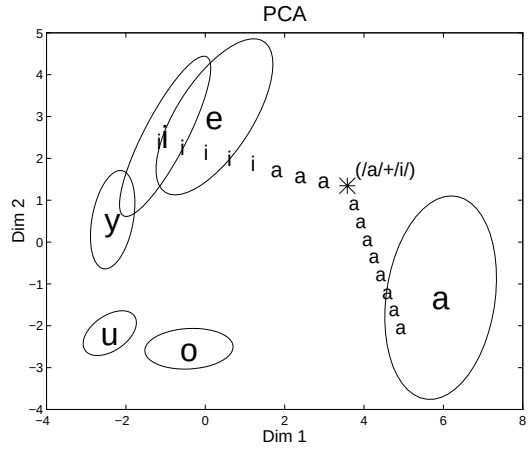

Figure 2: Segregation based on the discriminant analysis. Within a set of 6 vowels, segregation of the mixture /a/+/i/ shown by projection in the first PCA plane (Principal Component Analysis). The mixture (*) is identified as the dominant vowel /a/, and the iterative subtraction of the prototype of this class enables the identification of the second vowel /i/.

double-vowels is about 80%. Following the same principle, we have also developed a double-identification method using the Multi Layer Perceptron (**MLP**) [6]. MLP decomposition is not complete because the goal is to perform identification of members of the pair: the constraint is to have two different labels, and the consequence is to only get a *just-sufficient* variation of the spectra. By looking at the spectra at the end of the double-labelling process, we have observed that a linear representation is more appropriate to ground a decomposition when summation is quasi-linear.

4. EVALUATION OF THE MIXTURE COEFFICIENT

Input is known to be a weighted sum of two different spectra: $x=\alpha s_1+\beta s_2$. This is restricted to one degree of freedom, because of the normalisation of x: $\beta=1-\alpha$.

The scalar value can be estimated if both classes C_1 and C_2 of vectors s_1 and s_2 are given. We substitute the two unknown spectra by the renormalised mean spectrum of each class C_1 and C_2:

$$x = \hat{\alpha}\hat{s}_1 + \hat{\beta}\hat{s}_2, \text{ with } \|\hat{s}_1\| = 1 \text{ and } \|\hat{s}_2\| = 1$$

Then, we form the system with scalar products between input mixture and mean spectra:

$$(1) \quad \begin{cases} x.\hat{s}_1 = \hat{\alpha} + \hat{\beta}\hat{s}_1.\hat{s}_2 \\ x.\hat{s}_2 = \hat{\alpha}\hat{s}_1.\hat{s}_2 + \hat{\beta} \end{cases}$$

By solving this linear system, we find the mixture coefficient estimate:

$$(2) \quad \hat{\alpha} = \frac{\det(M_1)}{\det(M)}$$

$$\text{with } M = \begin{pmatrix} 1 & \hat{s}_1.\hat{s}_2 \\ \hat{s}_1.\hat{s}_2 & 1 \end{pmatrix} \text{ and } M_1 = \begin{pmatrix} x.\hat{s}_1 & \hat{s}_1.\hat{s}_2 \\ x.\hat{s}_2 & 1 \end{pmatrix}$$

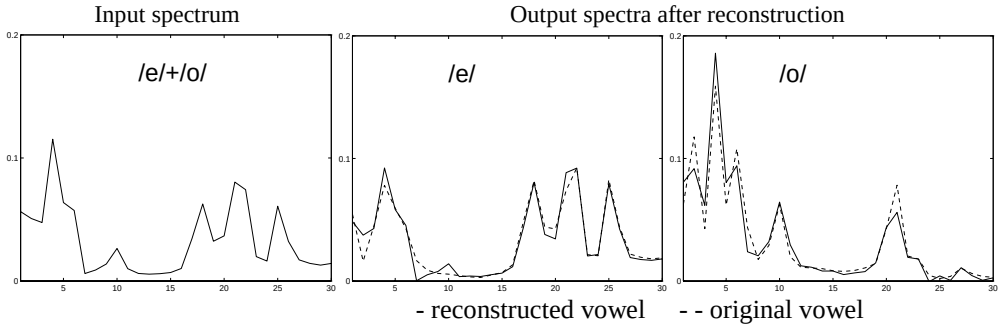

Figure 3: Reconstruction of the double vowel /e/+/o/ (temporal summation at 0dB RMS).
/e/ is the dominant vowel. Evaluated mixture coefficient = 0.89. Linear correlation with original vowel is 0.97 for /e/ and 0.96 for /o/

Here, the mixture coefficient is a scalar value varying between 0 and 1. We assume this is constant across the (30) spectral dimensions. This is an approximation: when summation occurs in the temporal domain, the relative phase spectra is expected to disrupt *homogeneity* of mixture coefficient vector across the frequency domain. Hence, the previous method is not able to compensate, and we will evaluate the consequence of this approximation by comparing *temporal* summation of signals and *spectral* summation. A second point concerns the potential properties of the vectorial form, expected to well support **(1)** decomposition of inhomogeneous summations (not due to previous effect), **(2)** partial overlapping, **(3)** subband processing [8] and **(4)** integration of multimodal information. For example, if the input vector is the spectrum, representing the mixture, appended with "clean" input values only representing source 1, mixture coefficients of supplementary channels are set to 1. Finally, mixture coefficient evaluation can be re-iterated during a second pass, by introducing in Eq. (1) the spectra obtained after reconstruction.

5. THE RECONSTRUCTION METHOD

Assuming that distribution of class C is a multidimensional Gaussian, and knowing the variance/covariance matrix, we can compute for a given spectrum x the probability $P(x|C)$ that C contains x, from the Mahalanobis distance. Having the labels (e.g., knowing classes C_1 and C_2) of the components of the input sum x, and the mixture coefficient, we can retrieve the spectra x_1 and x_2 such as $x = \alpha x_1 + \beta x_2$. We maximise the probability $P(x_1|C_1)*P(x_2|C_2)$. This corresponds to a minimisation of the sum $(d_1 + d_2)$ of the Mahalanobis distances defined by:

$$(3) \qquad \begin{cases} d_1 = (x_1 - \mu_1)^T M_1^{-1} (x_1 - \mu_1) \\ d_2 = (x_2 - \mu_2)^T M_2^{-1} (x_2 - \mu_2) \end{cases}$$

where μ_1, M_1 and μ_2, M_2 are respectively the mean and variance/covariance matrix of C_1 and C_2. According to the maximum likelihood method, $(d_1 + d_2)$ reaches a minimum when the first derivative is zero:

$$(4) \qquad \begin{cases} \dfrac{\mathrm{d}(d_1 + d_2)}{\mathrm{d}\, x_1} = 0 \\ x = \alpha x_1 + \beta x_2 \end{cases}$$

Developing (4), we easily obtain the following system:

$$(5) \qquad \begin{cases} x_1 = \left(M_1^{-1} + \dfrac{\alpha^2}{\beta^2} M_2^{-1} \right)^{-1} \left(M_1^{-1} \mu_1 - \dfrac{\alpha}{\beta} M_2^{-1} \mu_2 \right) \\ \qquad\qquad + \left(\beta^2 M_1^{-1} + \alpha^2 M_2^{-1} \right)^{-1} \alpha M_2^{-1} x \\ x_2 = \dfrac{x - \alpha x_1}{\beta} \end{cases}$$

Eq. (5) gives directly the two reconstructed spectra. Remark that, to compute variance/covariance matrices M_i, spectra are normalised by summation and removal of the mean. In order to perform matrix inversion without error precision, we have reconditioned the matrices M_i. After testing, we decided that addition of a slight Gaussian noise in the data before computing the variance/covariance matrices M_i is sufficient to get significant non zero values of $\det(M_i)$.

6. SIMULATIONS

6.1 Procedure

We have tested the reconstruction method (see [6] for the evaluation of the double-identification method described in 3) on a database of French synthetic vowels (/a/e/i/o/u/y/), with fundamental frequency (F0) varying between 100 and 200Hz (10Hz steps), and with frequency of the two first formants F1/F2 randomly varying according to the natural vowels variation (10 ex. each). The signal is Hamming-windowed and high-pass filtered. Then, a Fourier transform is applied and the amplitude spectrum is warped between 100 and 5000Hz according to the Mel scale (e.g., an auditory like representation). The test set is synthesised in the same manner, but with F0 varying between 105 and 195Hz, and other set of F1/F2 randomisation. Finally, the input data set includes 3000 pairs randomly selected in the test database, from the 13500 possible pairs.

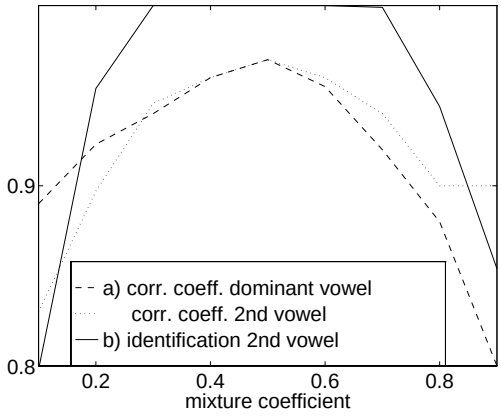

Figure 4: Effect of a biased estimation of the mixture coefficient α. Effective value $\alpha = 0.5$, with spectral summation and heuristic H2. Reconstruction uses α varying from 0.1 to 0.9. a) Correlation coefficient with the original vowels. b) Identification rate of the second vowel.

The principle is to test separately the influence of the three steps. To test the *reconstruction* ability of steps (2) and (3) without influence of step (1), we evaluate correlation coefficient between the original spectrum and the reconstructed one, when *both labels are given* to the system (H1, Table 1, column 2). The second index is the percentage of second vowels correctly labelled with the *late identification* heuristic (H2, Table 1, column 3). Only pairs in which the dominant label is found by step (1) (discriminant analysis) are *selected* for this statistic (the same procedure is used for Figure 4). To test step (2), we compare results when mixture coefficient is given or evaluated.

6.2 Results

Correlation coefficients and recognition rates (of the second vowel) depend on both relative level and condition of summation. Mean correlation coefficients are very high in all conditions, proving the ability of the method to well reconstruct both spectra (a typical example is shown Figure 3). Let us detail the results. First, when the mixture coefficient is *balanced*, at 0.5, with spectral summation, recognition rates of the second vowel are 100% when the first vowel is correctly identified. When the mixture coefficient is 0.8, we observe a small decrease of performance, only when the coefficient is evaluated. The effect of temporal summation is a decrease of about 10%. But part of this decrease (about 5%) is due to the 0dB RMS temporal summation, allowing an unbalanced spectral summation with average mixture coefficient at 0.8. Similar results are obtained either this is given or evaluated. Thus, decrease in "given" condition can be due to a bias between average and effective value. But we conclude it is partly (about 5%) due to inhomogeneity in both cases. The residue (5%) could be produced by the unbalanced condition. Figure 4 shows the effect of a bias on correlation coefficient and recognition rate of the second vowel. We observe that identification of the second vowel is still robust between 0.3 and 0.7, allowing a great *tolerance* in balanced condition.

Mixture type (Relative level)	Mixture coefficient α	H1 Correlation coefficient 1st - 2nd vowel	H2 Identification rate of the 2nd vowel
Spectral ($\alpha = 0.5$)	Given	0.97 - 0.97	100%
	Evaluated	0.95 - 0.95	100%
Spectral ($\alpha = 0.8$)	Given	0.99 - 0.93	99.4%
	Evaluated	0.98 - 0.92	94.4%
Temporal (0dB RMS)	Given: 0.8	0.95 - 0.92	89%
	Evaluated	0.96 - 0.92	90.6%

Table 1: Results of reconstruction and identification with spectral and temporal conditions of summation.

7. CONCLUSION

The three steps of the model are schema-based, but we have mentioned this model can be coupled with a primitive segregation stage. A primitive stage is expected to perform segregation by its own, partially or completely (as in [4]). Thus, the first *entry point* is to give to the schema-based level at least the label of the *dominant* vowel (and optionally its corresponding spectrum). The second entry point to couple levels is the mixture coefficient evaluation. Primitive stage could convey: **(1)** pointers on "clean" dimensions to do partial recognition [9] **(2)** confidence measure for attributes [7] **(3)** directly the vector of mixture coefficients. In this way, we have previously shown that such primitive stage outputs are evaluated after normalised binaural cross-correlation [10].

Finally, our purpose was not to present this decomposition model as biologically plausible, but this is a very suggestive one to represent "cortical resonance" to known stimuli. Object segregation is probably the result of *non linear* interactions occurring in the brain at the neural network level. We show that a linear model is sufficient to segregate analytically linear mixtures. Hence, decomposition of non linear mixtures is expected to appeal non linear and iterative procedures.

8. REFERENCES

[1] Varga, A.P. & Moore, R.K., ICASSP-90, 845-848, 1990.
[2] Bregman, A.S., ASA, MIT Press, 1990.
[3] Scheffers, M.T., Ph.D thesis, Groningen Univ., the Netherlands, 1983.
[4] Berthommier, F. & Meyer, G., this proc., 1997.
[5] Rose, R.C., Hofsterrer, E.M. & Reynolds, D.A., IEEE Trans. On Speech and Audio Proc., 2:2:245-257, 1994.
[6] Varin, L. & Berthommier, F., in proc. of the ESCA workshop, Keele, 222-225, 1996.
[7] Gaillard, F., Berthommier, F., Feng, G. & Schwartz, J-L., this proc., 1997.
[8] Bourlard, B. & Dupont, S., in proc. of Int. Conf. on Spoken Lang. Proc., Philadelphia, 422-425, 1996.
[9] Cooke, M., Morris, A & Green, P., in proc. of the ESCA workshop, Keele, 297-300, 1996.
[10] Tessier, E., Berthommier, F. & Meyer, G., in proc. of the 4th CFA, Marseille, vol. 1, 495-498, 1997.