IMPROVING OF AMPLITUDE MODULATION MAPS FOR F0-DEPENDENT SEGREGATION OF HARMONIC SOUNDS

Frédéric BERTHOMMIER* & Georg MEYER+

*Institut de la Communication Parlée/INPG
46, Av. Félix Viallet
38031 Grenoble CEDEX, FRANCE
bertho@icp.grenet.fr

+Department of Computer Science
Keele University
Keele, Staffs. ST5BG, UK
georg@cs.keele.ac.uk

ABSTRACT

The AM-map model [1] can be improved by adding two supplementary integration stages: the pooled map and the identification map. The pooled map's representation corresponds to a systematic bottom-up grouping of the first harmonics extracted at the level of the primary AM map. The identification map's representation corresponds to a classification of spectra segregated along the pitch axis. This labelling allows selection at the pooled map level of the two salient vowels according to the distribution of energy across the pitch axis. The selected labels are those associated with the higher peaks. During this selection stage, F0s are not given. Simulations show that the model is able to separate spectra according to F0 differences. The model therefore predicts qualitatively **(1)** the ability of listeners to segregate concurrent vowels, and **(2)** the effects of vowels' duration and relative level on segregation performance.

1. INTRODUCTION

We have shown previously [1] that the amplitude modulation map (**AM map**) is an intermediate representation of complex sounds allowing separability of spectral cues according to the fundamental frequency (F0). In order to solve the F0 tracking problem, a new model is proposed here that combines both bottom-up (or primitive) and top-down labelling information. The decision stage, which is used to select salient spectra, works at the intermediate level. The segregation power of the whole model is mainly primitive because only a simple schema-based segregation based on *subtraction* is engaged when the primitive process fails.

First, the F0-dependent recovering of spectrum is re-considered, and we add a supplementary representation after the primary AM map, which is the pooled map (**pAM map**). In order to validate the primitive segregation stage Assmann and Summerfield [2] proposed to perform harmonic grouping with the *auto-correlation*, applied on each channel of the peripheral representation (the cochlea and the auditory nerve).

We have previously shown that AM map is able to replace auto-correlation. The principle is a *demodulation* of the signal, channel by channel, followed by a Fourier decomposition. The identification process was preceded by a F0-guided segregation of spectral components (e.g., formants). F0s were assumed to be identified separately, and 'given' to the model. Now, the pooled map is a good *substitute* to autocorrelation. Both auditory models take into account the temporal coding of F0 performed at the peripheral level and are apparently able to predict relatively well the segregation performance of human listeners. It is thus necessary to differentiate between these two models. One method is to assess to segregation power of each model using the classical double-vowel segregation paradigm and looking at the effects of changing the *relative levels* of each vowel.

2. THE NEW MODEL

Main characteristics of the new model are included in Fig.1. Three stages correspond to the primary AM map model, connected to the pooled map and the identification map.

2.1 The primary AM map (AM map)

Stationary complex signals, such as vowels, are added (Input) and processed through a gammatone filterbank and weighted by an audiogram (Pre-proc. stage, Fig. 1). To build the primary AM map, a FFT is computed channel by channel after *demodulation* (rectification and bandpass filtering in the pitch domain). The representation is *bi-dimensional*: the first axis is the spectral one, and the second axis represents the envelope spectrum, related to *periodicity* information. The output of the filterbank is coded temporally and recoded spectrally by the FFT analysis. The modulation *envelope* is produced by the beating of *unresolved* harmonics in the medium and high frequency domain. Because the signal envelopes are not pure cosines, the AM spectrum contains harmonics.

With this model, the second axis allows a *representation* of the fundamental frequency in each channel. Energy is distributed along this axis because harmonics are produced. A supplementary process is necessary to pool (e.g., to group) the harmonics. This is implemented with by an harmonic sieve. The goal is to *recover* the spectrum of each signal presented. The pooling process is limited to the first five to ten harmonics whereas at least twenty harmonics have to be selected in Parsons' model [3] of segregation by harmonic selection. We select harmonics which are resolved harmonics in the low frequency domain and those of the AM spectrum, appearing after demodulation, in the high frequency domain.

Given precise F0 estimates, the model causes minimal *overlap* between competing sounds and hence minimal *distortion* of the recovered spectra. This previous version of the model, as all other existing segregation models, is F0-guided so that a parallel estimation of F0 is needed.

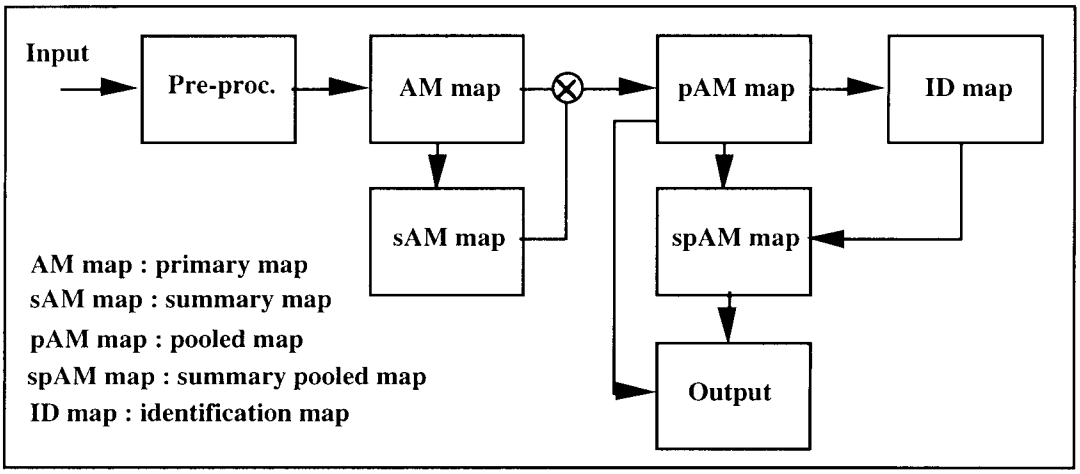

Figure 1: B*lock-Diagram of the model. sAM map and spAM map are summary maps having only the pitch dimension and result of a summation of energy across channels of spectral axis.*

It is effective when F0 differences are as small as one FFT bin (5-10Hz). Recognition scores obtained with segregated spectra can consequently be very good but depend critically on the *precision* of the F0 estimation stage. With the autocorrelation method, which is the main alternative to the AM map, F0 estimation is obtained by selecting the peaks in the summary representation [2]. In this method, energy of harmonic peaks is localised along the delay axis so that it is not necessary to perform a supplementary grouping process. Selection of candidates in the summary map is direct. Similarly, we estimated the pitch in a summary AM map (**sAM map**) using a harmonic *sieve* method.

The main advantageous properties of primary AM map are: (**1**) to allow *resolution* according to the *window length*, (**2**) to be a *quasi-linear* representation. Introduction of a supplementary pooling process intrinsically performed by autocorrelation is needed.

2.2 The pooled map (pAM map)

We propose to pool directly by *summing* multiple 'partial' spectra that are distributed relatively to the F0 axis in the primary map.

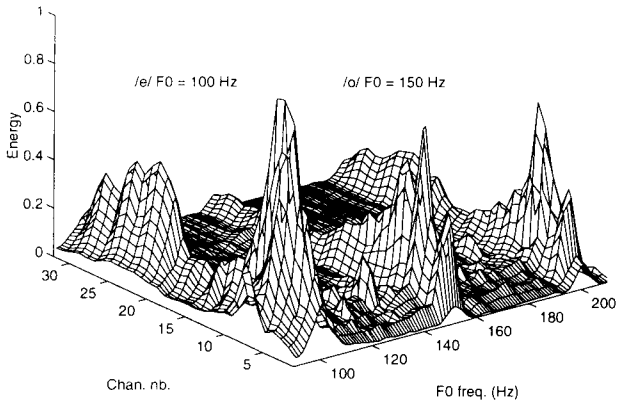

Figure 2: *Representation of /e/ 100Hz +/o/ 150 Hz (0dB, 200ms) in the pooled map. We can see the first harmonic of /e/.*

The summation of five to ten harmonics is preceded by an interpolation of five to ten, and completed by weighting with sAM map. This is in order to *eliminate* spurious peaks caused by the systematic summation, when no peak already exists in sAM map. For example, when 100/200/300Hz are present in the AMmap, a peak at 150Hz will appear, resulting from the summation of existing energy at 300Hz. This is eliminated by multiplication with sAM map, because there is no energy at 150Hz. Notice that sAM map takes into account peaks allowed by demodulation in the high frequency domain as well as the fundamental peak. After pooling, we evalute the square root. The pooled map is a two-dimensional representation of sounds where each point relates energy to frequency and modulation frequency. Peaks are resolved well (an example is shown Fig. 2).

We show that segregation of complex sounds becomes simply F0-dependent. Explicit (external) F0 identification is not necessary because each peak appearing in the new map corresponds to a candidate signal at a pitch that is given by the position on the map. A first version of the selection process aims to identify the highest peaks of the map, but we remark that this selection process is not satisfactory. There are at least three effects which will *complicate* a simple peak-picking strategy to segregate two candidates: (**1**) with close F0s, when the two peaks are not well resolved (**2**) the second peak in amplitude is often a harmonic of the first (**3**) the first harmonic of the second candidate is often unmasked, but its amplitude is lowered.

2.3 Coupling with an identification level

To complete our model, we propose to apply the identification process *systematically* to each spectral frame along the F0 axis, in order to build a high-level identification map (the **ID map**, Fig. 1). Building of the ID map consists in a crosscorrelation between normalised inputs (frames along the pitch axis) and every stored prototype, which are also normalised. The most correlated prototype is selected to label the summary pooled map (**spAM map**), so that a label is associated

with each pitch value (top-down arrow from ID map to spAM map, Fig. 1). This allows us to solve the problem of detection of secondary peaks and shoulders we mentioned before, by adding a supplementary information with a pattern matching stage. A typical example is shown Fig. 3, taking the same mixture as for Fig. 2.

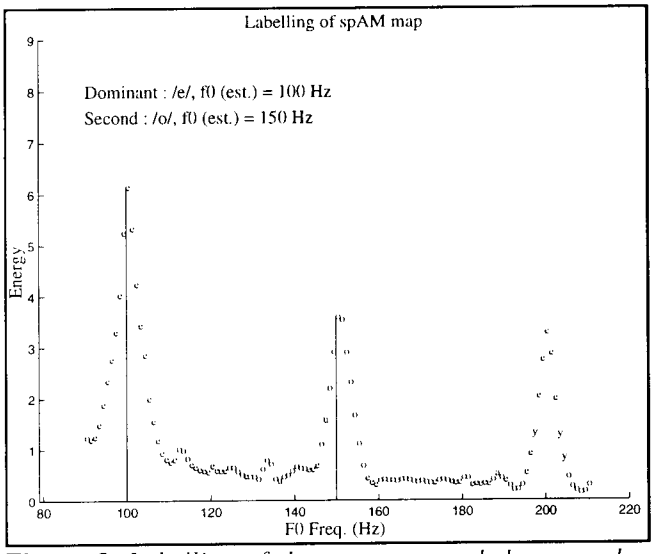

Figure 3: *Labelling of the summary pooled map and identification of /e/ 100Hz +/o/ 150Hz (0dB, 200ms). This is completed by pitch estimation.*

The most important implication of this method is that the selection is no longer driven by an explicit pitch estimation (thereby violating the principle of minimal commitment). With this representation, the pitch of the recognised signal is an emergent property of the segregation process as has been suggested by Bregman [4]. F0 identification can be a secondary product of this selection process, because when we select a peak in the upper-level, it is directly related with a F0 place (est. F0, Fig. 3).

Concurrent signals are segregated on F0 by virtue of the representation, not because of an explicit extraction process, which would require an external pitch estimate. The principle of *minimal commitment* is upheld because decisions on how to interpret the complex picture is deferred to the pattern matching stage. We propose to label energy before decision. The double vowel identification task consists in selecting in spAM map the two higher levels of energy associated with two different labels, and to report these two labels (Output may also be the two spectra, Fig. 1). If only one label is detected a *subtraction* process is applied to the main peak to obtain the masked spectrum and to label it. Hence, this case occurs more frequently when input signals have identical or close F0s.

2.4 The subtraction algorithm

The subtraction process we have implemented is a *decomposition* of an input spectrum V (a vector which is the frame of the main peak) by reference to previous prototypic spectra. Practically, these are centroïds of classes. Supposing that V is a weighted sum of unknown prototypes (for ex. V = a1 V1 + a2 V2), we evaluate label and contribution of the *dominant* prototype, which is the closest one to the mixture and we remove it in order to identify the residue. This needs an estimation of the first weight by the way of computation of scalar products between V and the n normalised prototypes, here:

V.V1 = a1 + a2 V2.V1 + ...
V.V2 = a1 V1.V2 + a2 + ...
V.V3 = a1 V1.V3 + a2 V2.V3 + ...
.
with Vi.Vi = 1, because the normalisation

This forms a system of n equations having n unknown weighting coefficients ai. We summarise by taking the *correlation* matrix M between all prototypes: the vector <V.Vi> of scalar products is linearly related to a vector of weights <ai> by <V.Vi> = M <ai>, so that this system gets solution when det(M) not nul. Because only the weight of the dominant member is needed, the *maximum* V.Vj of scalar products V.Vi is determined to get its label j. The first weight is aj = det(Mj)/det(M), with Mj obtained by substitution in M of the column j by the column vector <V.Vi>. The contribution of the dominant (for ex., j=1) is *subtracted* from V so that the scalar product between residue and any of the n prototypes Vi becomes:

((V - a1 V1).Vi) = V.Vi - a1 V1.Vi = 0 + a2 V2.Vi + ...

Hence, the weight associated with j=1 becomes 0. The second label is found by getting the maximum of this new vector of scalar products. By generalisation to mixture of multiple elements, this leads to the so called Iterative Linear Separation (**ILS**), which is a simple decomposition algorithm. We *iterate* the selection of dominant and the evaluation of weight in the residue.

3. DOUBLE VOWEL SEGREGATION

A main property of AM map representation is to allow simple control of the binding between spectral features with F0s attributes. The interpretation of the intermediate representation can be the simple extraction of a spectral frame for a given (or selected) F0 followed by a pattern matching stage. In our previous studies, F0s were given rather than estimated to study the segregation performance of the AM maps independently from any error introduced by pitch tracking [5]. This model was able to *predict* human performance for a range of experimental conditions. A pitch estimation stage was expected to degrade recognition scores by about 10%. The main features of this procedure are retained because the new representation is an extension of the primary AM map representation, which produces similar spectral outputs. The pooled map is based on the first 5 harmonics of pitches between 90 and 210Hz, with 1-Hz resolution. Identification rates are computed by cross-correlation of all input frames, extracted along the F0 axis, with references templates obtained by driving the model with isolated vowels.

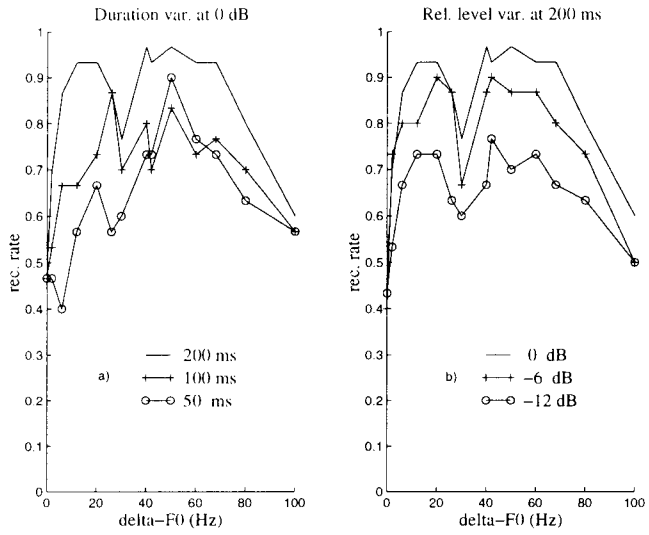

Figure 4: *Double recognition rates with variation of duration (4a, left) and relative level (4b, right).*

Recognition rates for both constituent parts of concurrent vowel pairs are shown Fig. 4. Pairs of stationary synthetic vowels within /a,e,i,o,y,u/ have a first member with F0=100Hz and a second one with F0 varying between 100 and 200Hz. The duration of the stimuli has three levels: 50, 100, and 200ms. The *rms level* of the second member also varies with three levels: 0, -6 and -12 dB. Evaluation is performed with all possible non-identical pairs of the six vowels. Figure 4a shows a variation of performance with signal duration for a range of F0 values of the second vowel. The 3 curves have a maximum at about 150Hz, indicating a clear delta-F0 dependence of segregation in the 3 cases. We need 12Hz difference to reach the plateau for 200ms, 26Hz for 100ms and 50Hz for 50ms: larger F0 differences are necessary to segregate vowels with shorter duration. The time-dependence of scores is the consequence of the resolution allowed by the Fourier transform. Autocorrelation models do not show this effect, although it is seen in human data. Performance also deteriorates with relative level difference (Fig. 4b). The reduction in performance is broadly similar, but the flattening of the curve is more pronounced at -12 dB than with 50 ms. This may indicate that the delta-F0 cue tends to deteriorate as the relative level differences become large.

The design of these simulations (linked with psychophysical experiments done in parallel) has been done in order to evaluate segregation power when information content decreases, by shortening the window of time and by decreasing the relative level. We show in Table 1 that the matrix of average performances is rather *symmetric*. Decays expressed relatively to the best performances have similar amplitudes along rows and columns. For example, we can compare a -12dB rel. level (the rms ratio is 4) - 64% (19% decay) - with a quarter shortening in time at 0dB (200-50 ms) - 63% (20% decay). This is globally consistent with the quasi-linear characteristics of our model and with the basic property of the Fourier transform (e.g., the time-frequency resolution trade-off). The amount of masking of the dominant element of a pair over the other one decreases with integration time because of an increasing resolution. Symmetrically, it varies directly with the relative amplitude of the masker in a linear model.

Time/rl	200 ms	100 ms	50 ms	
0 dB	83 (0)	70 (13)	63 (20)	**70 (0)**
-6 dB	76 (7)	61 (22)	56 (27)	**61 (9)**
-12 dB	64 (19)	52 (31)	47 (37)	**54 (16)**
	74 (0)	**61 (13)**	**55 (19)**	**6 3**

Table 1: *Average recognition performances (%) over delta-F0 with co-variation of time and rel. intensity levels. Decay rel. to (0dB, 200ms), in parenthesis.*

Finally, recovering of F0s is also satisfactory, and both F0 are recovered in a range depending on time window (about 10Hz at 100ms). When F0 does not correspond (e.g., it is not in a small range) with any F0 of the presented pair (or with a harmonic), this generally leads to a labelling error.

4. CONCLUSION

The improved AM-map is able to process vowel pairs without F0 tracking. This suggests that natural double-vowels can be successfully segregated, even when F0s are difficult to identify precisely because of interferences. However, overall performance still depends critically on the duration/stationarity of the signal and on the robustness of the classification stage. Segregation performance for additive natural vowels remains to be evaluated. The improved AM-map model is an updated tool providing a representation of sounds that accounts for both place and time coding. It is useful for **(1)** understanding auditory processes in terms of representation of sounds along perceptual dimensions such as pitch and timbre, and **(2)** evaluating the relative importance of primitive and schema-based segregation properties in vowel segregation (streaming is mainly performed at the primitive stage level in the present model). Further study of the effects of the duration and relative level of the concurrent vowels will complete the *comparison* between the autocorrelation [6] and the AM-map model. We expect that the autocorrelation model will be more sensitive than the AM-map model to changes in the relative level of concurrent vowels, and less sensitive than the AM-map model to changes in vowels duration.

Acknowledgements : Thanks to Christian Lorenzi for reading and comments on this paper.

5. REFERENCES

[1] F. Berthommier and G.F. Meyer (1995), in Proc. Eurospeech Madrid, pp. 135-138.
[2] P.F. Assmann and Q. Summerfield (1990), JASA, 88:2, 680-697.
[3] T.W. Parsons (1976), JASA, 60:911-918.
[4] A.S. Bregman (1990), ASA, MIT Press, London.
[5] G.F. Meyer and F. Berthommier (1996), in Proc. of Workshop on the Aud. basis of speech perc., Keele, pp. 212-215.
[6] G.F. Meyer and F. Berthommier (1997), Brit. J. Audiol., 31:108-110.