SPARSE NON-NEGATIVE DECOMPOSITION OF SPEECH POWER SPECTRA FOR FORMANT TRACKING

Jean-Louis Durrieu, Jean-Philippe Thiran

Signal Processing Laboratory (LTS5), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Many works on speech processing have dealt with auto-regressive (AR) models for spectral envelope and formant frequency estimation, mostly focusing on the estimation of the AR parameters. However, it is also interesting to be able to directly estimate the formant frequencies, or equivalently the poles of the AR filter. To tackle this issue, we propose in this paper to decompose the signal onto several bases, one for each formant, taking advantage of recent works on nonnegative matrix factorization (NMF) for the estimation stage, further refined by sparsity and smoothness penalties. The results are encouraging, and the proposed system provides formant tracks which seem robust enough to be used in different applications such as phonetic analysis, emotion detection or as visual cue for computer-aided pronunciation training applications. The model can also be extended to deal with multiple-speaker signals.

Index Terms— Speech Analysis, Autoregressive (AR) Model, Source-Filter Model, Non-negative Matrix Factorization, Sparse Decomposition.

1. INTRODUCTION

The source/filter model from [1] is the generally admitted model for speech or singing voice signals. The advantages of such a model lie in its link with the physical production process underlying the signal. Many speech processing techniques rely on its assumptions in order to estimate the fundamental frequency (F0) [2] or the spectral envelope, for instance by means of the Mel-Frequency Cepstral Coefficients (MFCC) for speech recognition [3].

The formant frequencies, *i.e.* the vocal tract resonance frequencies f_p (with p the formant number), correspond in this model to the poles of the filter part. This filter can be modeled as an all-pole filter, leading to an AR model for the speech signal. In order to estimate these frequencies or the resulting spectral envelope, many works have focussed on the estimation of the AR coefficients, as in [4], or through linear predictive coding (LPC) or peak-picking analysis, after a first estimation of the smooth spectral envelope [5, 6, 7]. Such approaches suffer from several disadvantages. The AR parameters are not linearly related to the poles and the estimation of the spectral envelope and the corresponding formant frequencies are done separately: it is therefore difficult to impose constraints such as the smoothness of formant tracks during the envelope estimation. A spectral envelope estimated this way may not exhibit the desired poles, at expected formant positions, hence the need for various post-processing steps. Furthermore, these methods usually assume that the signal consists of only one speaker, and can hardly be extended to deal with multiple-speaker signals.

We propose to estimate the spectral envelope of speech signals, within a framework allowing the control of the smoothness of the f_p tracks during the estimation process. The power spectrum of the signal is decomposed onto several bases, the elements of which are parameterized either by their f_0 (source basis) or by their f_p (individual formant basis). These bases are redundant, and a strategy aiming at imposing some sparsity is introduced to obtain a meaningful decomposition. The desired track smoothness is then built on top of this sparsity penalty strategy.

The resulting formant tracks on a vowel database are close to the desired tracks, although some hand-tuning of the parameters still seems necessary. The results can already be used for applications such as phonetics analysis [8], emotion detection [9] or as a visual support for computer-aided pronunciation training [10]. Although the scope of this paper is limited to the single-speaker case, another advantage of the proposed model is the possibility to extend it to signals with several speakers, as suggested by previous works using a similar model [2]. This may for instance be used for speaker separation [11].

This paper is organized as follows. The signal model is first discussed. Then the parameter estimation algorithm is derived, along with the sparsity and smoothness strategies. The results are then presented and we conclude with perspectives for this work.

2. SIGNAL MODEL

In this section, the model for one frame is first described, and then extended to consecutive frames.

2.1. Model for one signal frame

The input signal x is an utterance by a speaker, which is assumed to follow a source/filter model [1]. The filter part is assumed to follow an all-pole model. Let z_p, $p = 1 \ldots P$, be its distinct poles, *i.e.* the real and complex conjugate pairs of poles, where P is the number of formants. Let s_f be the power spectrum at frequency bin f such that $s_f = |x_f|^2$, with x_f the Fourier transform of size F of a frame from x. The expression of s_f, for complex poles z_p, is given by:

$$s_f = \prod_{p=1}^{P} \underbrace{\left|(1 - z_p e^{-2j\pi f/F})(1 - z_p^* e^{-2j\pi f/F})\right|^{-2}}_{s_f^{F_p}} s_f^{F_0} \qquad (1)$$

where $\mathbf{s}^{F_p} = [s_0^{F_p}, \ldots, s_F^{F_p}]^T$ is the squared magnitude of the frequency response of the AR(2) filter, with poles $\{z_p, z_p^*\}$, and $\mathbf{s}^{F_0} = [s_0^{F_0}, \ldots, s_F^{F_0}]^T$ the power spectrum of the excitation (or source) signal. When the latter is voiced, it is mainly parameterized by the

This work was funded by the Swiss CTI agency, project n. 11359.1 PFES-ES, in collaboration with SpeedLingua SA, Lausanne, Switzerland.

The authors would like to thank G. Richard and B. David for the discussions which led to this article and S. Arberet for his comments on the article.

fundamental frequency f_0, hence the super-script F_0. For a pole $z_p = \rho_p \exp 2j\pi f_p/F_s$, where F_s is the sampling rate of x, f_p is referred to as the p^{th} formant frequency.

The proposed framework enables to jointly estimate the spectral envelope $\prod_{p=1}^{P} s_f^{F_p}$, the pitch f_0 and the formant frequencies f_p. Contrary to previous works, using LPC for instance, the proposed framework aims at allowing formant frequency tracking directly during the envelope estimation process. It can also be used for multiple speaker mixtures. Instead of estimating the f_p values, for each p, we generate several AR(2) filter frequency responses (for $p = 1 \dots P$) with different values of (f_p, ρ_p) and several glottal source power spectra (for $p = 0$) with different f_0. Each $\mathbf{s}^{F_p}$ is then approximated as a sparse decomposition on its corresponding power spectrum dictionary, similarly to [12].

In practice, for $p \in [0, P]$, the vector $\mathbf{s}^{F_p}$ is modeled as a non-negative linear combination of K^p fixed power spectra $\mathbf{w}_k^p$, stored as the column vectors of the $F \times K^p$ matrix $\mathbf{W}^p = [\mathbf{w}_0^p, \dots, \mathbf{w}_{K^p}^p]$. Therefore:

$$s_f^{F_p} = \sum_{k=0}^{K^p-1} w_{fk}^p h_k^p = [\mathbf{W}^p \mathbf{h}^p]_f, \tag{2}$$

where $\mathbf{h}^p$ is the vector of the decomposition coefficients. Eq. (1) can therefore be written with matrix conventions, with $\bullet$ the Hadamard product:

$$\mathbf{s} = (\mathbf{W}^1\mathbf{h}^1) \bullet \dots \bullet (\mathbf{W}^P\mathbf{h}^P) \bullet (\mathbf{W}^0\mathbf{h}^0) \tag{3}$$

For the **filter part**, *i.e.* $p \in [1, P]$, each column of $\mathbf{W}^p$ is the frequency response of an AR(2) filter, with complex poles $\{z_{kp}, z_{kp}^*\}$. The formant frequency range for each p is set such that $f_{kp} \in [F_p^{\min}, F_p^{\max}]$. As in [12], a grid of linearly spaced values for f_{kp} and $\rho_{kp} \in \{0.8, 0.84, 0.88, 0.92, 0.97\}$ is used, with $G_F = 60$ values for f_{kp} and $G_R = 5$ values for ρ_{kp}. Let $g_F \in [0, G_F - 1]$ and $g_R \in [0, G_R - 1]$, then $k = G_R g_F + g_R$ and, with $\lfloor y \rfloor$ the integer part of y:

$$f_{kp} = F_p^{\min} + (F_p^{\max} - F_p^{\min})\lfloor k/G_R \rfloor/(G_F - 1)$$

With such a choice, elements in $\mathbf{W}^p$ that share the same frequency f_{kp} have consecutive indices k. This ordering strategy in $\mathbf{W}^p$ is important for the subsequent sparsity and smoothness-inducing procedures. Elements from a dictionary $\mathbf{W}^p$ are shown on Fig. 1. The following table summarizes the chosen formant ranges:

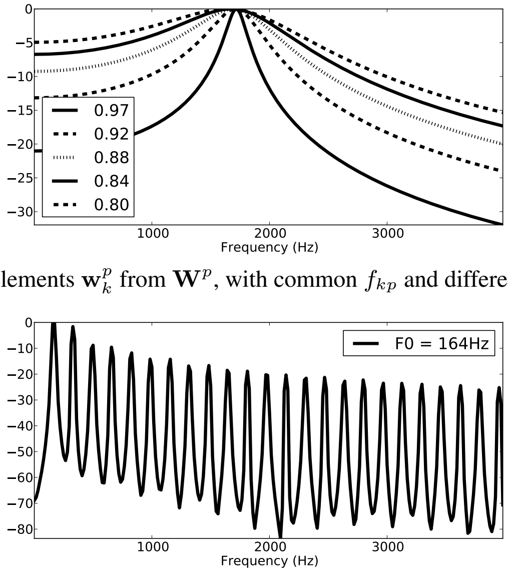
p	$F_p^{\min}$	$F_p^{\max}$	comments
1	200	1000	from [13, 8]
2	550	3100	-
3	1700	3800	-
4	2400	6000	additional values
5	4500	8000	-

We are mostly interested in the first three formants: P should therefore be greater than 3. In [8] an LPC of order 14 was used for the estimation, corresponding to a maximum of $P = 7$ distinct complex poles. We chose $P = 5$, because greater values with the proposed framework tend to provide inconsistent results, with duplicated formants.

$\mathbf{W}^0 = [\mathbf{w}_0^0, \dots, \mathbf{w}_U^0]$ is the basis for the **source part**: each column $\mathbf{w}_u^0$ is generated as the power spectrum of a glottal signal whose F0 is:

$$\mathcal{F}(u) = F_0^{\min} \times 2^{\frac{u-1}{12U_{\text{st}}}}, \tag{4}$$

where $F_0^{\min}$ is the lowest desired F0, and U_{st} the number of F0 per semitone. Setting a maximum value $F_0^{\max}$ for F0 therefore sets U, the number of elements in $\mathbf{W}^0$. The values used for this study are $F_0^{\min} = 80$Hz, $F_0^{\max} = 500$Hz and $U_{\text{st}} = 16$. The glottal source model KLGLOTT88 [14] was used, with an open coefficient arbitrarily fixed to 0.6. The order in which the columns are stored in $\mathbf{W}^0$ also allows for an easy smoothing (or tracking) procedure.



(a) Elements $\mathbf{w}_k^p$ from $\mathbf{W}^p$, with common f_{kp} and different ρ_{kp}.

(b) An element $\mathbf{w}_k^0$ from $\mathbf{W}^0$.

Fig. 1. Elements from the dictionaries $\mathbf{W}^p$ (values in dB).

For each basis $\mathbf{W}^p$, an extra column of constant value ($w_f = 1, \forall f$) is appended. For the source part, it can be seen as the unvoiced part of the speech production, while for the filter part, it mainly allows to discard a formant, notably when the given range does not fit the actual data.

The signal frame is assumed to be generated by one F0, and one formant frequency for each formant p. In principle, this means that the decomposition in Eq. (3) should be **sparse**, with only one non-null coefficient in vector $\mathbf{h}^p$. It is not computationally feasible to consider all the possible combinations, and select the best-fitting one. An approximate solution is proposed: the vectors $\mathbf{h}^p$ are first estimated without constraint. Within the iterative algorithm developed in Section 3, the estimated $\mathbf{h}^p$ are then **re-weighted** at each iteration, gradually enforcing the non-null coefficients to concentrate in one region, hence simulating an $\mathcal{L}^0$ sparsity penalty, a technique comparable with that of [15].

2.2. Modeling consecutive frames

We are in general interested in tracking the F0 and formant frequency contours, *i.e.* determining at each frame n of the signal the underlying frequencies f_{pn}, $\forall p$. The power spectra $\mathbf{s}_n$, $\forall n = 1 \dots N$, are stacked as the columns of a matrix $\mathbf{S} = |\mathbf{X}|^2$, the power of the short-term Fourier transform (STFT) of x. Using Eq. (3), the model for $\mathbf{S}$ writes:

$$\mathbf{s}_n = (\mathbf{W}^1\mathbf{h}_n^1) \bullet \dots \bullet (\mathbf{W}^P\mathbf{h}_n^P) \bullet (\mathbf{W}^0\mathbf{h}_n^0) \tag{5}$$

$$\mathbf{S} = (\mathbf{W}^1\mathbf{H}^1) \bullet \dots \bullet (\mathbf{W}^P\mathbf{H}^P) \bullet (\mathbf{W}^0\mathbf{H}^0), \tag{6}$$

where the link between the proposed model and non-negative matrix factorization (NMF) becomes obvious. This parallel is useful when designing the estimation algorithm of Sec. 3.

The desired F0 and formant frequencies also need to be constrained to “slowly” evolve. In [2], this **smoothness** is imposed *a posteriori* for the F0 sequence. It is here proposed to impose it during the estimation process, with the same re-weighting strategy as for the sparsity constraint, as discussed in Section 3.3.

3. PARAMETER ESTIMATION

3.1. Maximum Likelihood (ML) criterion

The parameters of model (6) are estimated by ML. Each element x_{fn} of the STFT $\mathbf{X}$ is a complex Gaussian variable, centered, with a variance equal to s_{fn}, defined in Eq. (6). Estimating $\boldsymbol{\Theta} = \{\mathbf{H}^p\}_{p=0\ldots P}$ in this ML framework is equivalent to finding the set $\widehat{\boldsymbol{\Theta}}$ which minimizes the Itakura-Saito (IS) divergence between $\mathbf{X}$ and $\mathbf{S}$ (which depends on $\boldsymbol{\Theta}$), *i.e.* the following criterion $C(\boldsymbol{\Theta})$ [16]:

$$C(\boldsymbol{\Theta}) = \sum_{fn} -\log \frac{|x_{fn}|^2}{\prod_p \left(\sum_k w^p_{fk} h^p_{kn}\right)} + \frac{|x_{fn}|^2}{\prod_p \left(\sum_k w^p_{fk} h^p_{kn}\right)} \tag{7}$$

To avoid the scale indeterminacies of criterion $C(\boldsymbol{\Theta})$, $\forall p = 1\ldots P$, $\sum_k h^p_{kn} = 1$. $\mathbf{H}^0$ therefore bears most of the energy of the signal.

3.2. Estimation algorithm

As in [2], a multiplicative gradient approach is used to minimize C. For $h^p_{kn} \in \boldsymbol{\Theta}$, the derivative is of the form:

$$\nabla_{h^p_{kn}} C = \nabla^+_{h^p_{kn}} - \nabla^-_{h^p_{kn}} \tag{8}$$

where $\nabla^+_{h^p_{kn}} \geq 0$ and $\nabla^-_{h^p_{kn}} \geq 0$. Then, in practice, updating h^p_{kn} the following way decreases the value of C [17]: $h^p_{kn} \leftarrow h^p_{kn} \times \nabla^-_{h^p_{kn}} / \nabla^+_{h^p_{kn}}$

The estimation algorithm then consists in iterative updates of the $P+1$ matrices, as follows:

1. Initialize $\mathbf{H}^p$, $\forall p$, with positive random values,
2. Repeat for a given number of iterations $I = 100$:
   • For p from 0 to P, compute $\mathbf{S}^{F_p}$ and $\mathbf{S}$, then update $\mathbf{H}^p$:

$$\mathbf{S}^{F_p} = \mathbf{W}^p\mathbf{H}^p, \quad \mathbf{S} = \mathbf{S}^{F_0} \bullet \mathbf{S}^{F_1} \bullet \ldots \bullet \mathbf{S}^{F_P}$$

$$\mathbf{H}^p \leftarrow \mathbf{H}^p \bullet \frac{\boldsymbol{\nabla}^-}{\boldsymbol{\nabla}^+} \text{ with } \begin{cases} \boldsymbol{\nabla}^- = (\mathbf{W}^p)^T \dfrac{|\mathbf{X}|^2}{\mathbf{S}^{F_p} \bullet \mathbf{S}} \\ \boldsymbol{\nabla}^+ = (\mathbf{W}^p)^T \dfrac{1}{\mathbf{S}^{F_p}} \end{cases}$$

   where the operators $\bullet$ and $(.)^2$ as well as $\div$ are element-wise.

3.3. Sparsity and regularity of the coefficients

The desired estimations should be sparse, allowing to identify one formant or F0 frequency f_p per frame, and these frequency contours should be smooth, since the modifications of the vocal tract cannot be instantaneous. A method for imposing the sparsity is first introduced, leading to a convenient way of enforcing the smoothness.

Sparsity: For a given p, at a given frame n, the vector $\mathbf{h}^p_n$ is assumed to have its energy mostly concentrated around a single component at $k = \mu^p_n$. Classical $\mathcal{L}^1$ or $\mathcal{L}_2$ sparse estimations as in [17] are not structured, and therefore do not comply with the desired concentration of energy or hardly scale to the smoothing problem which is later discussed. Instead, a heuristic **re-weighting strategy** is preferred. It first consists, after each new estimate of $\mathbf{h}^p_n$, in finding an index k at which the energy in $\mathbf{h}^p_n$ is concentrated. h^p_{kn} is then reinforced against the other values in $\mathbf{h}^p_n$, with degrees depending on the iteration number.

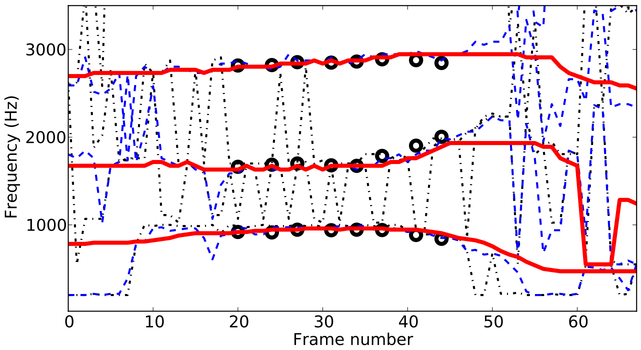


Fig. 2. Formant track estimates for file 'b13ah'. Circles: GT tracks, dot-dashed: unconstrained, dashed: sparse, plain: smooth estimates.

At each iteration i of the above algorithm, for $p = 1\ldots P$, the “mode” μ^p_n is computed as the barycenter of the index k of $\mathbf{h}^p_n$ (weighted by its values). For $p = 0$, a term giving more importance to lower frequencies is included in the weights to avoid octave errors:

$$\mu^0_n = \sum_k \frac{h^0_{kn}(K^0 - k)^2}{\sum_l h^0_{ln}(K^0 - l)^2} k \text{ and } \mu^p_n = \sum_k \frac{h^p_{kn}}{\sum_l h^p_{ln}} k \tag{9}$$

The weighting function is chosen as a Gaussian function, centered at $k = \mu^p_n$, with variance σ_i, $\forall n, p$. The σ^p_i are chosen as a decreasing function, hence narrowing the lobe over the I iterations. $\mathbf{h}^p_n$ is then re-evaluated as:

$$h^p_{kn} \leftarrow \omega^p_{kn} h^p_{kn}, \text{ with } \omega^p_{kn} = \exp - \frac{(k - \mu^p_n)^2}{2\sigma^2_i} \tag{10}$$

The maximum of ω^p_{kn} is equal to 1 and corresponds to the mode, when $k = \mu^p_n$. With decreasing values of σ_i, the number of non-null coefficients decreases, down to 1, which is the desired sparsity in the proposed framework. The chosen sequence of σ_i is exponentially decreasing, from K^p down to 3, in order to allow some flexibility in the model.

In the proposed dictionaries, whose elements are ordered according to their formant or F0 frequencies, neighboring elements have close values of f_p. Allowing more than one element around the mode can then be seen as allowing for some variation around f_p. This motivates the use of a final $\sigma_i > 1$, compensating potential misfit between the model and the data.

Smoothness: At last, given the above modification to take the desired sparsity into account, it is straightforward to smooth the f_p tracks by constraining, for each p and at each iteration, the sequence $\{\mu^p_n\}_n$ to be smooth. Note that, contrary to [17], the smoothness of the sequence of indices k is sought after, and not the smoothness of the values of $\{h^p_{kn}\}_n$. For this work, after the computation of each mode, a median filter smoothes the sequence. Thanks to the chosen ordering in $\mathbf{W}^p$, $\forall p$, smoothing the sequence of index also implies the smoothness of the sequence of f_p frequencies (as well as ρ_p sequences for $p > 0$). An example of the estimated formant tracks is given on Fig. 2.

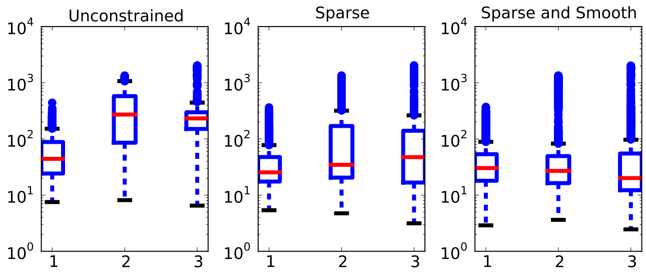


Fig. 3. “Box and Whiskers” plot of $\{\epsilon_{ppv}\}_{p=1,2,3}$, for each system. Box: lower to upper quartiles; line: median; dots: fliers.

4. RESULTS

4.1. Vowel database and evaluation criterion

The proposed algorithm was evaluated on the vowel database of [8], consisting of 1668 `/hVd/` utterances, with `V` a vowel from the “phonetic” set {`ae`, `ah`, `aw`, `eh`, `er`, `ei`, `ih`, `iy`, `oa`, `oo`, `uh`, `uw`}. The sampling rate is 16 kHz. The STFTs are computed on 32ms-long windows, every 8ms, weighted by a Hann window. For each utterance, the first 3 formants are annotated on 8 “reliable” locations.

Three “systems” are compared, based on the proposed model: the “unconstrained” estimation, the “sparse” estimation and the “smooth” estimation (which includes sparsity).

We compare each “ground-truth” (GT) formant track $\{f^0_{qn}\}_n$ to each estimated formant track $\{f_{pn}\}_n$. The tracks are compared only on the 8 frames for which the annotation is available. For each file v, the deviation between the different estimated tracks $\{f^v_{pn}\}_n$ and GT tracks $\{f^{0,v}_{qn}\}_n$ are computed as the mean squared error, in Mel: $\epsilon^2_{qpv} = \frac{1}{8}\sum_n |f^v_{pn} - f^{0,v}_{qn}|^2$.

4.2. Discussions

As shown on Fig. 2, the estimated smooth tracks closely follow the annotated tracks. The unconstrained and the sparse estimates also cluster around the GT tracks, but exhibit chaotic sequences. The sparsity penalty is needed to obtain smooth meaningful sequences, but it does not provide such sequences by itself. This is confirmed by the distribution of errors between the estimated f^v_{pn} and the corresponding GT $f^{0,v}_{pn}$, as shown on Fig. 3.

The distributions of the errors (*i.e.* the histograms of $\{\epsilon_{qpv}\}_v$), for the “smooth” system, are such that there is a salient peak near 0 only when $p = q$: the first three GT formants can most of the time be identified with the first three smoothed formants. This is mostly true for the utterances by the “men” and “women” groups, but for the “boys” and “girls” group, some confusions occur more often, associating formant f_2 to f^0_1 and f_3 to f^0_2. To overcome this issue, a de-correlation penalty between formant f_1 and f_2 could be added.

The “smooth” formant performance, in terms of quartiles, is close to that of an **LPC analysis** of order 14 (as was used for the annotation), with median errors of about 20 Mels. Although the proposed method is not as efficient as the LPC, its advantages are the possibility to estimate the **poles** and to enforce both their **sparsity**, and the **smoothness** of their tracks, along with the potential extension to **multiple-speaker** signals analysis.

Note at last that the estimation of the F0 pitch track is also performed. The performance with respect to this feature was not evaluated, but the effectiveness of this type of decomposition was shown in [2] for singing voice signals.

5. CONCLUSION AND PERSPECTIVES

We proposed a new approach for the estimation of formant tracks, based on an AR(P) speech model, inspired by NMF decompositions of power spectra and refined by sparsity and smoothness control during the estimation stage.

The smoothed estimated tracks are consistently related to the GT formant tracks of the database from [8]. Other penalties can be included using the estimation/re-weighting framework: for instance, de-correlation penalties or weights reflecting some other *prior* or conditional knowledge may be used.

Future studies should assess the performance of the proposed method on longer utterances such as phrases or sentences. The resulting features, the pitch and the formant tracks, can be used in several other applications, such as phonetics, speaker separation, but also for emotion detection or as a visual feedback within computer-aided pronunciation training softwares.

6. REFERENCES

[1] G. Fant, *Acoustic Theory of Speech Production*, Mouton, 1970.

[2] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, “Source/filter model for unsupervised main melody extraction from polyphonic audio signals,” *IEEE TASLP*, vol. 18, no. 3, pp. 564 –575, March 2010.

[3] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE TASSP*, vol. 28, no. 4, pp. 357 – 366, Aug. 1980.

[4] R. Badeau and B. David, “Weighted maximum likelihood autoregressive and moving average spectrum modeling,” in *IEEE ICASSP*, 2008, pp. 3761–3764.

[5] Paul Boersma, “Praat, a system for doing phonetics by computer,” *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, 2001.

[6] L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals*, Prentice-Hall signal processing series. Prentice-Hall, 1978.

[7] F. Villavicencio, A. Robel, and X. Rodet, “Improving Lpc Spectral Envelope Extraction Of Voiced Speech By True-Envelope Estimation,” in *proc. of the IEEE ICASSP*, 2006.

[8] J.M. Hillenbrand, L.A. Getty, M.J. Clark, and K. Wheeler, “Acoustic characteristics of American English vowels,” *JASA*, vol. 97, pp. 3099–3111, 1995.

[9] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, “Feartype recognition for future audio-based surveillance systems,” *Speech Comm.*, vol. 50, no. 6, pp. 487–503, 2008.

[10] A. Cazade, “De l’usage des courbes sonores et autres supports graphiques pour aider l’apprenant en langues,” *ALSIC*, vol. 2, no. 2, pp. 3–32, 1999, *in French*.

[11] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, “Single channel speech and background segregation through harmonic-temporal clustering,” in *WASPAA*, 2007, pp. 279–282.

[12] N. Moal and J.J. Fuchs, “Estimation de l’ordre et identification des paramètres d’un processus ARMA,” in *GRETSI*, 1997, *in French*.

[13] R.W. Schafer and L.R. Rabiner, “System for automatic formant analysis of voiced speech,” *JASA*, vol. 47, pp. 634, 1970.

[14] D.H. Klatt and L.C. Klatt, “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” *JASA*, vol. 87, pp. 820–857, 1990.

[15] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, “A fast approach for overcomplete sparse decomposition based on smoothed $\mathcal{L}^0$ norm,” *IEEE TSP*, vol. 57, no. 1, pp. 289–301, 2009.

[16] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural Comp.*, vol. 21, no. 3, March 2009.

[17] T. Virtanen, “Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria,” *IEEE TASLP*, vol. 15, no. 3, pp. 1066–1074, 2007.