PARAMETRIC DECOMPOSITION OF THE SPECTRAL ENVELOPE

Anderson Fraiha Machado[1], *Antonio Bonafonte*[2], *Marcelo Queiroz*[1]

[1]Computer Science Department, University of São Paulo, Brazil,
[2]Universitat Politècnica de Catalunya, Barcelona – Spain

ABSTRACT

In this paper we propose a novel representation of the spectral envelope of speech, using sums of general parametric basis functions (filter bank), including Gaussian, Hann and Nutall. The central frequency and bandwidth of each filter can be adjusted independently for each speech frame as in conventional filter bank analysis. The quality of the representation has been evaluated with respect to a reference spectral envelope obtained from the Harmonic-Stochastic Model [1] (HSM), and also with respect to parameter time stability and separability of different acoustic classes. Numerical results show that the use of the basis functions proposed is an improvement over pure Gaussian models of the spectral envelope.

Index Terms— spectral envelope representation, spectral decomposition, parametric functions, speech modeling

1. INTRODUCTION

The estimation of the spectral envelope is an important issue in describing the speech signal production system. In many cases, the envelope must be represented using few parameters, either for achieving a high compression rate or for statistical modeling. On the other hand, when the speech signal is going to be reconstructed from the envelope, as is the case in statistical speech synthesis and voice conversion, it is crucial that the estimated envelope properly represents the vocal tract.

Whereas LPC is the most extensively used method to compute the spectral envelope of speech signals, Mel-Frequency Cepstrum Coefficients (MFCC) are the most widely-used spectral envelope representation for speech classification tasks (eg.: speech recognition, speaker verification, etc.); MFCCs can also be used in reconstructing speech with good results [2]. An alternative model was introduced in [3], where the spectral envelope is represented as a sum of few Gaussian basis functions, which are meant to have a good correspondence with spectral formant regions.

The representation of spectral envelope in terms of basis functions has some advantages over classical models, such as LPC or MFCC. For instance, it allows independent instantaneous access and modification of each spectral frequency band, and the simplified use of nonlinear frequency models, such as perceptual scales. It further allows a time-varying control of each spectral region, without seriously degrading the quality of the reconstructed speech signal.

In this paper we analyze different basis functions (filter shapes) for representing an input speech signal. By imposing different constraints in the estimation algorithm it is possible to offer a controllable trade-off between accuracy and efficiency of the method. In Section 2 we formulate the representation problem and describe the different basis functions considered. Several estimation methods to estimate the filters are presented in Section 2. Constraints on central frequency and bandwidth can be imposed to get different envelope estimations. Finally, in Section 3 the different filter shapes (basis functions) and estimation methods are evaluated using several objective functions: spectral distortion, stability of temporal evolution and acoustical inter and intra-class distribution of the parameters.

2. GENERAL BASIS FUNCTION DECOMPOSITION

The representation of spectral envelope using sums of parametric functions has been studied in recent literature [4]. The sum of L general parametric functions is defined as

$$\hat{S}(f; a, \mu, \sigma) = \sum_{k=1}^{L} \psi(f; a_k, \mu_k, \sigma_k) \qquad (1)$$

where each component $\psi(f; a_k, \mu_k, \sigma_k) : \mathbb{R} \longrightarrow \mathbb{R}$ corresponds to a continuous radial basis function with amplitude a_k, central frequency μ_k and bandwidth σ_k, evaluated at frequency f, normalized to lie between $-\pi$ and π.

Familiar examples of radial basis functions are the Hann, Nuttall and Gaussian window functions, among others. A particular class of basis functions based on cosine decomposition was used in this work, which are defined as

$$\psi(f; a_k, \mu_k, \sigma_k) = a_k \sum_{m=0}^{M} d_m \cos(2\pi m \frac{f - \mu_k}{\sigma_k}) \qquad (2)$$

if $f \in [\mu_k - \sigma_k, \mu_k + \sigma_k]$, and $\psi(f; a_k, \mu_k, \sigma_k) = 0$ otherwise. This formulation allows us to select several window-

This research was sponsored by CAPES and FAPESP grant 2008/08632-8. and by the Spanish Government under grants TEC2009-14094-C04-01 and TEC2012-38939-C03-02

shaped functions for the fitting, according to the choice of the parameters M and $\{d_m\}_{m=1}^{M}$, as can be seen in table 1.

Table 1. *Some coefficients d_k of each respective basis function.*

Function	d_0	d_1	d_3	d_4
Hann	0.5	−0.5		
Nuttall	0.35577	−0.48740	0.14423	−0.01260
Blackman-Harris	0.35875	−0.48829	0.14128	−0.01168
Blackman-Nuttall	0.36358	−0.48918	0.13660	−0.01064

In this work we have also used, for the sake of comparison, other basis functions such as Gaussian window. Two important requirements for other general basis functions to be used in this setting, which are true for the above examples, is that $\lim_{f \longrightarrow \mu_k \pm \sigma_k} \psi(f; a_k, \mu_k, \sigma_k) = 0$, so that the sum of all components is continuous, and that they are differentiable at all points with respect to the parameters $\theta = (a, \mu, \sigma)$, so that we can apply iterative estimation algorithms. Moreover, we consider that the partial derivatives with respect to a, μ, σ are computed within each bandwidth $[\mu_k - \sigma_k, \mu_k + \sigma_k]$; and use zero padding outside this frequency range in the case of bases originated from periodic functions.

The modeling problem based on radial basis functions is then to find an approximation $\hat{S}(f; a, \mu, \sigma)$ of a given discretized function $S(f)$ (i.e. known only at a finite set of frequencies) in such a way that the estimation error is minimal. In practice, for voiced frames, $S(f)$ corresponds to an unknown spectrum envelope, whose values can be estimated at harmonic frequencies $S(k\, f_0)$.

This approximation could in principle be found by any iterative method, such as Newton or steepest descent, that fits the sum of basis functions to a smooth version of $S(f)$ (e.g. a polynomial or spline approximation). However, it is not hard to encounter practical situations where decomposition parameters obtained by classical iterative methods do not change smoothly over successive voiced frames, even though the spectral envelope changes smoothly. We call such a representation temporally unstable. Many applications, as speech synthesis, require a temporally stable spectral representation, in which the parameters (a_k, μ_k, σ_k) of each basis function corresponding to a spectral subband vary slowly as functions of time. This work presents some proposals to tackle this problem, by restricting the positioning of each basis function $\psi(a_k, \mu_k, \sigma_k)$.

2.1. Initialization: Greedy Fitting

In many applications it is desired that parameters are estimated within fixed intervals, specially the means of each component which are usually confined to a set of given frequency bands. Consider a set of L frequency bands centered on $\{c_k\}_{k=1}^{L}$; in the case of sound spectra, the choice of the set of central frequencies $\{c_k\}$ might take into account our perceptual system. For instance, the set of central frequencies used in this work are uniformly distributed with respect to the Mel scale.

Given a set of central frequencies, we consider a subdivision of the original spectrum $S(f)$ in spectra $S_k(f)$, defined by windowing the original spectral signal S using a (window) function W_k centered on c_k. It is usual to assume that $\sum_k W_k \approx 1$, so that $\sum_k S_k = \sum_k S \cdot W_k \approx S$.

Notice that it is possible to define static parameters that are completely determined by S_k. For instance, by considering the global peak $p = (x, y)$ of each spectral sub-band S_k, we can define a_k^0 and μ_k^0 as y and x respectively, and set σ_k^0 so that the area bounded by S_k is the same as that below $\psi(f; a_k^0, \mu_k^0, \sigma_k^0)$. From this setting, it is possible to find the optimal amplitudes a_k^0 since the estimation error can be minimized using least squares optimization. This approach will be adressed as Greedy Algorithm in the Section 3. Also, setting up will be used as initialization step for the iterative approach presented below, which is based on Marquardt's algorithm.

2.2. Iterative Fitting Algorithm

The method described here is similar to Algorithm 1.1 in [5], which considered originally only sums of Gaussian components. It consists in a variant of Marquardt's algorithm [6] that include the addicional constraint $a_k > 0$ for all k. As opposed to [7], in which all parameters $(a_k, \mu_k, \sigma_k)\ \forall k$ are estimated simultaneously, we treat each basis function $\psi(a_k, \mu_k, \sigma_k)$ separately. The overall algorithm builds the sub-bands S_k and initializes the parameters for the basis functions according to the greedy strategy presented above 2.1.

The next step is to refine the component $\psi(a_k, \mu_k, \sigma_k)$ that approximates the spectral sub-band S_k using a `Base Fitting Algorithm` as follows. Each basis function is updated iteratively by making $(a'_k, \mu'_k, \sigma'_k) = (a_k, \mu_k, \sigma_k) + (\delta_{a,k}, \delta_{\mu,k}, \delta_{\sigma,k})$ and by choosing the parameter variations δ according to

$$[\mathbf{J}^T\mathbf{J}]\delta_k = \mathbf{J}^T[S_k(f) - \psi(f; a_k, \mu_k, \sigma_k)], \forall f \qquad (3)$$

where

$$\mathbf{J}(a_k, \mu_k, \sigma_k) = \left[\frac{\partial\psi}{\partial a}\frac{\partial\psi}{\partial\mu}\frac{\partial\psi}{\partial\sigma}\right](a_k, \mu_k, \sigma_k) \qquad (4)$$

is the Jacobian of $\psi(a_k, \mu_k, \sigma_k)$. This update corresponds to optimally fitting a linear model of $\psi(a_k, \mu_k, \sigma_k)$ to S_k in the parameter space (a, μ, σ). This iterative update is carried on until the approximation error, defined by

$$\mathcal{E}_k = \|S_k - \psi(a'_k, \mu'_k, \sigma'_k)\|^2 \qquad (5)$$

has a negligible variation between successive iterations, i.e. if $\Delta\mathcal{E}_k = |\mathcal{E}'_k - \mathcal{E}_k| \leq \epsilon$.

The above strategy can be adapted to produce sums of basis functions approximating $S(f)$ using fewer parameters, in

order to save representation space. This can be achieved by fixing some of the parameters μ and/or σ as global values, and optimizing the remaining parameters, using the reduced Jacobian matrix with respect to the free parameters in Equation 3. For instance, if μ is fixed ($\mu_k = c_k = \text{center}(W_k)$), then the reduced Jacobian is $\mathbf{J}(a_k, \sigma_k) = \left[\frac{\partial \psi}{\partial a} \frac{\partial \psi}{\partial \sigma}\right](a_k, \sigma_k)$, so that $\delta_k = (\delta_{a,k}, \delta_{\sigma,k})^T$ is given by

$$[\mathbf{J}(a_k, \sigma_k)^T \mathbf{J}(a_k, \sigma_k)]\delta_k = \mathbf{J}(a_k, \sigma_k)^T [S_k - \psi(a_k, \mu_k, \sigma_k)]. \quad (6)$$

Other expressions have to be modified accordingly, e.g. $\mathcal{E}_k = \|S_k - \psi(a'_k, \sigma'_k)\|^2$. This variation is called Free-$[a, \sigma]$, and one can define analogously the variants Free-$[a, \mu]$ and Free-$[a]$. For the sake of comparison with these variants, the original method will be referred to as Free-$[a, \mu, \sigma]$ in the experimental section. More details of this approach can be found in [5].

3. EVALUATION

Our aim in the following experimental evaluation is to show not only the accuracy of the spectral fitting obtained by the proposed method, but also the temporal stability of the parameters that represent each basis function component $\psi(a_k, \mu_k, \sigma_k)$. We also investigate some properties of clustering speech data based on the use of these parameters.

In all tests, the input is a voice signal composed of six segments with different vowels pronounced by the same speaker. The system segments the input signal in frames of 256 samples and obtains both harmonic and stochastic envelopes from the HSM model. In the case of LPC and cepstrum (CEPS) the harmonic and stochastic envelopes are estimated directly from the harmonic peaks and the log-module spectrum, respectively.

For the method here proposed, the following windows are used as parametrical bases: Hann, Nutthall, Blackman-Harris, Blackman-Nutthall and Gaussian. The methods used in the comparison are the Greedy Method, Free-$[a, \mu, \sigma]$, Free-$[a, \mu]$, Free-$[a, \sigma]$ and Free-$[a]$. The spectral envelopes obtained by these methods are compared to the spectral envelopes obtained from Cepstrum and LPC models of similar size. The threshold parameter used to stop the iterative fitting algorithm is $\epsilon = 10^{-5}$. The number of coefficients used in all methods is 24, with a sampling rate equal to 16 kHz.

The experimental design takes into account the following issues: the fitting accuracy measured as the distance between the reconstructed envelope and the true envelope; the stability of the dynamic temporal evolution of each model parameter; the distances between speech segments that belong to the same (artificial) phonetic class; and the distances between the centroids of these classes.

Comparisons are made frame by frame between the original signals s and reconstructed signals $\hat{s}$. The fitting accuracy considers the overall fitness of the spectral envelope for each proposed model with respect to a ground truth envelope, which is the harmonic envelope obtained in the HSM model; in this preliminary experiment we have used as fitting measure a normalized version of the Spectral Distortion (SD) between a given spectral envelope model and the ground truth. Consider that $S^{[n]}$ corresponds to the original spectrum of the n-th frame of the input signal s, with $n = 1, 2, \ldots, N$. Then, we can calculate the fitting accuracy as

$$E = \frac{1}{N} \sum_{n=1}^{N} \left\| S_{\log}^{[n]} - \hat{S}_{\log}^{[n]} \right\|^2 \quad (7)$$

where $S_{\log}$ is a normalized log-spectrum S defined as

$$S_{\log} = 10 \log_{10}(S + \varepsilon_0) \quad (8)$$

where ε_0 is a value that defines the normalization floor, which is in our case 10^{-7}.

The stability of the temporal evolution of each parameter is measured as the sum of distances between consecutive values of the parameters in each pair of adjacent frames. This measure basically is associated with discontinuity rate of signal along the time evolution. If $\mathbf{w}^{[n]} = (a^T, \mu^T, \sigma^T)^T$ is the vector of amplitudes, central frequencies and bandwidths in frame $\hat{S}^{[n]}$, then the temporal stability measure is

$$\Psi = \frac{1}{N-1} \sum_{n=2}^{N} \left\| \mathbf{w}^{[n]} - \mathbf{w}^{[n-1]} \right\|, \quad (9)$$

i.e. greater stability corresponds to smaller Ψ values.

Figure 1 presents the values of Spectral Distortion E and stability Ψ for each method.

Among the bases, a visible highlight to the Hann window is noticed. Figure 1 indicates that this choice combined with the Free-$[a, \mu, \sigma]$ method exhibits best fitness values, with low distortion spectral rates and good temporal stability. However, the Free-$[a, \mu]$ method is a better alternative than Free-$[a, \sigma]$ for representation using few coefficients. Although the Greedy method hasn't achieved highest scores, it stands out as the most efficient among the proposed methods, since it does not depend on iterative optimization procedures.

Another property expected from these representation models is their use as feature vectors describing the input signal, for instance in tasks such as clustering segments into (artificial) phonetic classes. Typically an artificial phonetic class uses the spectral centroid as its primary key. The measure called *within-class scatter* [8] for a class with centroid $\mathcal{C}_i$ is defined as

$$\gamma_i = \sum_{n=1}^{N_i} \left\| \mathbf{w}^{[n]} - \mathcal{C}_i \right\|^2. \quad (10)$$

In order to measure how far away are the phonetic classes from each other, we have used a variation of the *Fisher linear*

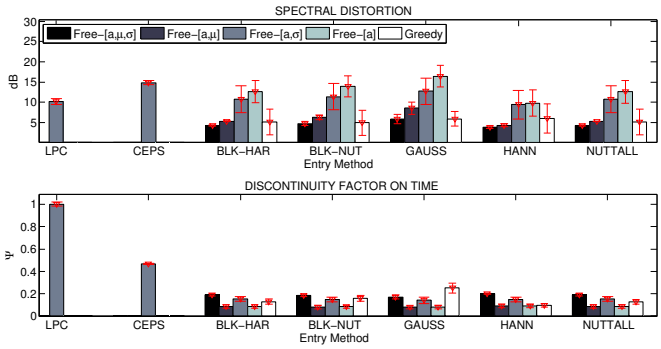

Fig. 1. Spectral Distortion and Discontinuity Rates

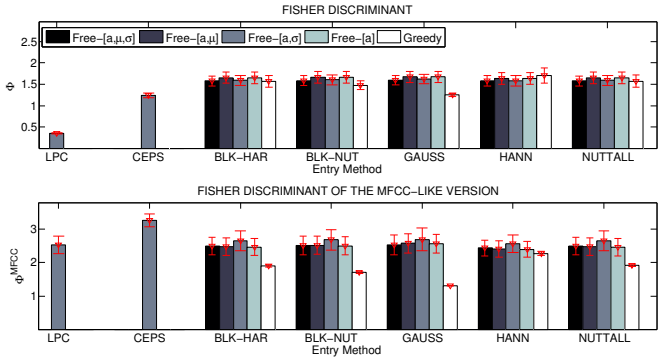

Fig. 2. Fisher Discriminant of spectral coefficients and Fisher Discriminant of MFCC-like versions

discriminant, which is defined as

$$\mathbf{\Phi} = \frac{\sum_{i=1}^{I} N_i \left\| \mathcal{C}_i - \mathbf{m} \right\|^2}{\sum_{j=1}^{I} \gamma_j}, \qquad (11)$$

where $\mathbf{m}$ is the global mean of all feature vectors of each phonetic class. In principle, we want to find models that provide small values of γ and large values of $\mathbf{\Phi}$.

In speech classification the MFCCs are frequently used. They offer a very good discrimination between sounds and speaker features. Figure 2 derives a MFCC-like representation based on the spectral estimation proposed in this paper. The amplitude values a_k (already sampled in log-scale) are multiplied by σ_k to get a factor closely related to energy values of standard MFCC. Finally, the DCT of these values is computed. The larger $\mathbf{\Phi}$ values of the transformed parameters confirm that these provide better discrimination than the original spectral parameters. According to bases selection criterion, the best discriminant features are obtained with Free-$[a, \sigma]$. However, these values should be validated with a larger and more general test.

4. CONCLUSION

In this paper we have presented an alternative representation method for the spectral envelope of speech signal segments, based on sums of general radial basis functions, and the methods required to compute such representations. Numerical experiments show that it is possible to achieve better fittings with respect to a reference spectral envelope than those obtained by LPC, MFCCs and sums of Gaussian components. These representations also achieve a better behavior in terms of temporal stability, with smaller parameter variation between successive frames of the speech signal.

Preminaries experiments show that the Hann method has presented the lowest distortion score. Regardless of the individual comparison among the proposed methods, we can conclude they are a good alternative for flexible modeling of the spectral envelope, with individual bank-filter control and good clustering properties.

Further work will include the application of these representation models in speech processing problems, such as pitch shifting and timbre modification (i.e. voice conversion).

5. REFERENCES

[1] D. Erro, A. Moreno, and A. Bonafonte, "Flexible harmonic/stochastic speech synthesis," in *6th ISCA Workshop on Speech Synthesis*, 2007.

[2] Dan Chazan, Ron Hoory, Gilad Cohen, and Meir Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2000.

[3] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of gaussians," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. IEEE, 1996, vol. 2, pp. 1229–1232.

[4] E. Godoy, O. Rosec, and T. Chonavel, "Speech spectral envelope estimation through explicit control of peak evolution in time," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE, 2010, pp. 209–212.

[5] Anderson Fraiha Machado, Antonio Bonafonte, and Marcelo Queiroz, "Spectral envelope representation using sums of gaussians," in *Iberspeech*, 2012.

[6] D.W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[7] Ardeshir Goshtasby and William D. O'Neill, "Curve fitting by a sum of gaussians," *CVGIP: Graphical Model and Image Processing*, vol. 56, no. 4, pp. 281–288, 1994.

[8] R.O. Duda, P.E. Hart, and D.G. Stork, "Pattern classification," *New York: John Wiley, Section*, vol. 1, pp. 654, 2001.