Spectral Envelope Representation using Sums of Gaussians

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

[1] Institute of Mathematics and Statistics, University of São Paulo, São Paulo – Brazil,
dandy@ime.usp.br, mqz@ime.usp.br,
[2] Universitat Politècnica de Catalunya, Barcelona – Spain,
antonio.bonafonte@upc.edu

Abstract. The aim of this paper is to present a new approach to spectral representation using sums of Gaussian distributions. Sums of Gaussians provide an intuitive representation for frequency bands of a signal spectrum as well as formant regions. The representation of spectral envelopes using Gaussian parameters $\{a, \mu, \sigma\}$ simplifies the expression of important tasks such as frequency warping and formant manipulation. Marquardt's algorithm has been extended to estimate parameters of the Gaussian models for each frequency band, allowing each Gaussian parameter to be either optimized for fitting a given spectral sub-band, or else have a fixed value for reducing the number of model parameters. This allows for several choices on the sets of free/fixed parameters and the sizes of models. Experimental results show that the models proposed offer an accurate approximation of spectral envelope, and provide good perceptual results when applied to pitch shifting.

Keywords: spectral envelope representation, sum of Gaussians

1 Introduction

The estimation of spectral envelopes is one of the most important and well-established topics in Signal Processing in general, and in Speech Technology in particular. Several well-known applications, such as speech recognition, speaker verification, statistical speech synthesis, diarization and voice conversion, require obtaining a faithful representation of the spectrum envelope using few parameters. The Mel-Frequency Cepstrum Coefficients (MFCC) are the most widely-used spectral envelope representation. In statistical speech synthesis LPC-based cepstrum coefficients are also frequently used. In speech recognition, we are interested in obtaining a representation which is speaker-independent, or at least one that can be easily transformed into a speaker-independent representation. For instance, vocal-tract length normalization (VTLN) has been successfully included in the adaptation of MFCC to reduce speaker variability and simplify speech recognition tasks. VTLN has also be used in voice conversion to transform the spectrum of a given acoustic signal from one speaker to another. Formant-based representations are an alternative which allows one to characterize a spectrum envelope using few parameters. However, this representation is not accurate enough and is not easily estimated under all conditions.

Gaussian-Mixture Models (GMM) are frequently used in speech technology to represent the probability density function (pdf) of speech parameters (e.g. MFCC of one particular speaker) [1]. Even if the pdf does not strictly fit a given model, it provides accurate representation for most practical situations, given that enough Gaussian components are included in the mixture. Well-known algorithms are widely-used to estimate the parameters of the pdf from training data (e.g: EM for maximum likelihood estimation). Many techniques have been proposed to adapt or transform the sum of Gaussians representing one acoustic signal class to another, (e.g: voice gender conversion [2] and voice conversion [3]).

In this paper we introduce a parametric spectral estimation method that assumes that the spectral envelope can be represented as a sum of Gaussians. Notice that the function to be represented – the spectral envelope – depends on only one dimension (the frequency) and is strictly positive. Moreover, it is expected that parameters of each Gaussian component (center frequency, amplitude and bandwidth) reflect temporal continuity between adjacent frames.

Several *models* are proposed in Section 2 with varying constraints, and estimation methods are derived. Section 3 presents an application framework for pitch shifting, which is a required step in statistical speech synthesis (and in some unit-selection speech-synthesis systems). Finally, Section 4 presents both objective and subjective evaluations of the quality of the transformed signal.

2 Gaussian Spectral Fitting

The problem of fitting a sum of Gaussians to a spectral representation of a signal has been addressed in previous work. In 1996, Zolfaghari [1] presented a method for formant estimation using mixtures of Gaussians based on the well-known EM Algorithm.Also, the same author has proposed a Bayesian modelling of the STRAIGHT spectral envelope [4] using mixtures of Gaussians [5]. Although these methods provide good results in representing formant regions, they are not able to model spectral nuances that are required in speech reconstruction.

A recent work [6] has also proposed a spectral envelope model using sum of Gaussians in which the amplitude and central frequency of each component is defined based on each local peak of the spectrum. The variance estimation considers the distance between neighboring peaks, and does not consider the overlaps between Gaussians in the mixture.

Unlike these methods, in what follows uses pre-defined frequency bands to get a optimal Gaussian parameters using a adaptation of Marquardt's algorithm.

The sum of L Gaussians is defined as $G(a, \mu, \sigma) = \sum_{k=1}^{L} G(a_k, \mu_k, \sigma_k)$ with each Gaussian $G(a_k, \mu_k, \sigma_k) : \mathbb{R} \longrightarrow \mathbb{R}$ given by

$$[G(a_k, \mu_k, \sigma_k)]\,(f) = a_k\, e^{-\frac{(f-\mu_k)^2}{2\sigma_k^2}}, \qquad (1)$$

where a_k, μ_k and σ_k are, respectively, the amplitude, central frequency and standard deviation of the k-th Gaussian component. The modelling problem is to find an approximation of a function $S(f)$ by the above sum of Gaussians in such a way that the estimation error is minimal.

In this work, $S(f)$ represents the envelope of the DFT of a discrete signal; for instance, $S(f)$ can be derived using a non-parametric representation of the spectral envelope. This fitting problem is equivalent to solving a system of nonlinear equations with $3L$ coefficients using a nonlinear least squares estimator. Furthermore, since the spectral envelope is assumed to have positive values, it becomes necessary to use an estimation method that constrains $a_k > 0$ for all k, which is addressed below.

2.1 Sequential Gaussian Fitting Algorithms

In 1994, Goshtasby and O'Neill [7] proposed a method to fit a curve with a sum of Gaussians based on the algorithm of Marquardt, whose initialization is determined from derivative values and zero-crossing rates. Although this proposed system obtains an accurate approximation,it is not applicable to the problem of spectral representation, since it allows the use of negative amplitudes.

This paper presents a variant of Marquardt's algorithm [8] constraining $a_k > 0$ for all k. The original proposal [7] estimates all Gaussian parameters (a_k, μ_k, σ_k) $\forall k$ in parallel, making it inappropriate for controlling specific parameters of each Gaussian, such as the value of a_k. Therefore we propose a modification where we treat separately each Gaussian component $G(a_k, \mu_k, \sigma_k)$, which fits a pre-specified section of the spectrum defined by a spectral sub-band S_k. There are two important steps of the proposed algorithm that are intertwined. The first one is the definition of the spectral sub-bands S_k, which is addressed below, and the second is the parameter estimation for each Gaussian $G(a_k, \mu_k, \sigma_k)$, so that it fits the respective spectral sub-band S_k, which is addressed afterwards.

Defining spectral sub-bands In some applications, particularly in speech processing (e.g. such as statistic synthesis [9] and voice conversion [3]), it is often desired that the Gaussian coefficients are estimated in fixed frequency bands, specially the means of each Gaussian component.

Given a set of L frequency bands centered on c_k, we first split the spectrum in L band-limited portions S_k, that will be fitted to a Gaussian component in the sum of Gaussians. Each S_k is defined by windowing the spectral signal with a window W_k, usually a triangular window, so that $\sum_k S_k = S$. The set of sub-bands used in this work are the critical bands of the Bark scale, represented as a set of windows B_k.

Algorithm 1.1 describes how each sub-band S_k is obtained and initially modelled by a static Gaussian, with its parameters obtained directly from S_k. The choice of initial Gaussian parameters uses the global peak $p = (x, y)$ of each spectral sub-band S_k, reestimated with parabolic interpolation; in this work, a_k^0, μ_k^0 and σ_k^0 are initialized as y, x and width of S_k, respectively.

Algorithm 1.1 $(a^*, \mu^*, \sigma^*) =$ `Band_Gauss_Estimation(`S`,`L`,`B`,`ϵ`)`

```
for k = 1 to $L$
    $S_k \leftarrow S \cdot B_k$
    $a_k^0 \leftarrow \max(S_k)$ /* using quadratic interpolation of amplitude */
    $\mu_k^0 \leftarrow \arg\max(S_k)$ /* using quadratic interpolation of frequency */
    $\sigma_k^0 \leftarrow$ bandwidth($S_k$)
    $(a_k^*, \mu_k^*, \sigma_k^*) \leftarrow$Gaussian_Fitting($S_k$,$(a_k^0, \mu_k^0, \sigma_k^0)$,$\epsilon$)
end for
```

Each Gaussian component $G(a_k, \mu_k, \sigma_k)$ is then optimally fitted to the spectral sub-band S_k using Algorithm 1.2 `Gaussian_Fitting`, which will be discussed next.

Optimally fitting each Gaussian component Given the k-th Gaussian, the goal of Algorithm 1.2 `Gaussian_Fitting` is at each iteration to update the parameters (a_k, μ_k, σ_k) by making $(a'_k, \mu'_k, \sigma'_k) = (a_k, \mu_k, \sigma_k) + (\delta_{a,k}, \delta_{\mu,k}, \delta_{\sigma,k})$, in such a way that $G(a'_k, \mu'_k, \sigma'_k)$ fits S_k better. The value $G(a'_k, \mu'_k, \sigma'_k)$ can be approximated by a linear function of $\delta_k = (\delta_{a,k}, \delta_{\mu,k}, \delta_{\sigma,k})^T$ as

$$G(a'_k, \mu'_k, \sigma'_k) \approx G(a_k, \mu_k, \sigma_k) + J(a_k, \mu_k, \sigma_k)\delta_k \tag{2}$$

where $J(a_k, \mu_k, \sigma_k) = \left[\frac{\partial G}{\partial a} \frac{\partial G}{\partial \mu} \frac{\partial G}{\partial \sigma}\right](a_k, \mu_k, \sigma_k)$ is the Jacobian of $G(a_k, \mu_k, \sigma_k)$ containing the derivatives of the Gaussian function with respect to its parameters:

$$\begin{aligned} \left[\frac{\partial G}{\partial a}\right](f) &= e^{-\frac{(f-\mu_k)^2}{2\sigma_k^2}} \\ \left[\frac{\partial G}{\partial \mu}\right](f) &= \frac{a_k(f-\mu_k)}{\sigma_k^2} e^{-\frac{(f-\mu_k)^2}{2\sigma_k^2}} \\ \left[\frac{\partial G}{\partial \sigma}\right](f) &= \frac{a_k(f-\mu_k)^2}{\sigma_k^3} e^{-\frac{(f-\mu_k)^2}{2\sigma_k^2}} \end{aligned} \tag{3}$$

The parameter δ_k that minimizes the error for the k-th Gaussian is given by

$$[J(a_k, \mu_k, \sigma_k)^T J(a_k, \mu_k, \sigma_k)]\delta_k = J(a_k, \mu_k, \sigma_k)^T [S_k - G(a_k, \mu_k, \sigma_k)] \tag{4}$$

where S_k is the k-th spectral sub-band and $J(a_k, \mu_k, \sigma_k)$ is the Jacobian of the k-th Gaussian $G(a_k, \mu_k, \sigma_k)$.

An important aspect of this algorithm is the stopping criterion in the search of a minimal solution, defined according to the rate of change of the absolute error $\mathcal{E}_k$ defined as

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

The estimation of each Gaussian ends when the variation of the error $\mathcal{E}_k$ between successive iterations is smaller than a given threshold, i.e. if $\Delta\mathcal{E}_k = |\mathcal{E}'_k - \mathcal{E}_k| \leq \epsilon$ is verified. The following algorithm estimates the parameters $(a_k^*, \mu_k^*, \sigma_k^*)$ of the k-th Gaussian that best fits the k-th spectral sub-band S_k. It depends on the threshold ϵ and on the initial values of the Gaussian parameters.

Algorithm 1.2 $(a_k^*, \mu_k^*, \sigma_k^*) =$ `Gaussian_Fitting(`S_k`,`(a^0, μ^0, σ^0)`,`ϵ`)`

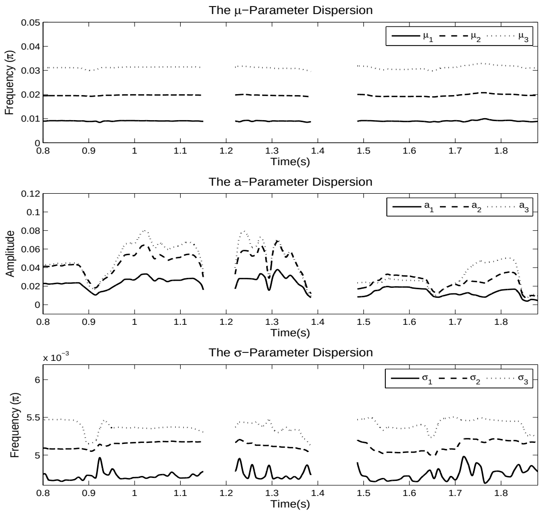

Fig. 1. *Dynamic Evolution of Gaussian Parameters.*

```
$\mathcal{E}_k^0 \leftarrow ||S_k - G(a_k^0, \mu_k^0, \sigma_k^0)||^2$
$\Delta\mathcal{E} \leftarrow \infty$
$(a_k, \mu_k, \sigma_k) \leftarrow (a_k^0, \mu_k^0, \sigma_k^0)$
while $\Delta\mathcal{E} > \epsilon$
    Calculate $J(a_k, \mu_k, \sigma_k)$ using Eq. 3
    Calculate $\delta_k$ by Eq. 4 using $J(a_k, \mu_k, \sigma_k)$
    $(a_k, \mu_k, \sigma_k) \leftarrow (a_k, \mu_k, \sigma_k) + \delta_k$
    Calculate $\mathcal{E}_k$ by Eq. 5
    $\Delta\mathcal{E} \leftarrow |\mathcal{E}_k^0 - \mathcal{E}_k|$
    $\mathcal{E}_k^0 \leftarrow \mathcal{E}_k$
end while
$(a_k^*, \mu_k^*, \sigma_k^*) \leftarrow (a_k, \mu_k, \sigma_k)$
```

Figure 1 displays the time evolution of the parameters $(a_k^*, \mu_k^*, \sigma_k^*)$ computed by Algorithm 1.1 on a sentence uttered by an arbitrary speaker. This evolution should be soft along the time, since the input sound is a vowel sustained /a/. In order to represent a set of L Gaussians, $3L$ parameters are used, which can become an expensive model depending on the application. The next section discusses alternative ways of representing the sum of Gaussians that keep some of these parameters fixed, making the size of the model smaller.

2.2 Parameter Constraints

An alternative use of sums of Gaussians to model spectral envelopes using fewer parameters consists of keeping some of the parameters μ and/or σ fixed as global values, and to optimize the remaining parameters with respect to fitting the corresponding spectral sub-bands.

In this case, given one or two fixed parameters, we have to adjust the Jacobian matrix to maintain just the remaining (free) parameters within its structure. For instance, if we consider that μ is fixed ($\mu_k = c_k = \text{center}(B_k)$), then the reduced Jacobian used by this method is defined as

$$J(a_k, \sigma_k) = \left[\frac{\partial G}{\partial a} \frac{\partial G}{\partial \sigma} \right] (a_k, \sigma_k) \tag{6}$$

Notice that the Jacobian matrix $J(a_k, \sigma_k)$ is a submatrix of the complete Jacobian defined in Equation 3, where the derivative with respect to μ has been eliminated. Other definitions have to be similarly adapted; for instance, in the above example $\delta_k = (\delta_{a,k}, \delta_{\sigma,k})^T$ and $\mathcal{E}_k = ||S_k - G(a'_k, \sigma'_k)||^2$. The optimal parameters will be given by a similarly modified version of Equation 4; considering μ fixed this modified equation will be:

$$[J(a_k, \sigma_k)^T J(a_k, \sigma_k)]\delta_k = J(a_k, \sigma_k)^T [S_k - G(a_k, \mu_k, \sigma_k)] \tag{7}$$

Everything else works exactly the same way. It should be easy to see how this method would work for σ fixed ($\sigma_k = \text{bandwidth}(B_k)$) and (a, μ) free.

A special case of parameter space reduction occurs when both μ and σ fixed. In this case the only free parameter would be a, and we would have $J(a_k) = \frac{\partial G}{\partial a}(a_k)$ and $\delta_k = \delta_{a,k}$ would have to verify

$$J(a_k)^T J(a_k)\delta_k = J(a_k)^T \left[S_k - G(a_k, \mu_k, \sigma_k)\right].$$

This special case where μ and σ are both fixed could also be addressed by a much more efficient method, where the amplitudes of all Gaussian components are estimated in one step, in order to minimize the overall error of the approximation $\|S - G(a, \mu, \sigma)\|$ where $G(a, \mu, \sigma) = \sum_{k=1}^{L} G(a_k, \mu_k, \sigma_k)$. If we maintain the restriction $a_k > 0$ for all k, this problem is reduced to a Non-Negative Least Squares Estimation - NNLSE [10].

Analogously, the same process of jointly estimating Gaussian amplitudes with μ and σ fixed can be used as a refinement step in each algorithm previously proposed. Obviously this implies in raising processing costs, as can be seen in Table 1. Here we present the Mean Square Error of each processed frame in an arbitrary signal, and also the number of iterations of the innermost loop, with and without the NNLSE refinement step. The conditions of this test will be described on Sec. 4 using the sawtooth synthetic signal.

It is seen in this table that all methods have a similar behavior, where the refinement step produces better solutions at the cost of increased computation.

3 Application on Pitch Shifting

In this section, we will discuss the application of the proposed models to pitch shift a voiced signal using the Harmonic plus Stochastic Model. A pitch shifting

Table 1. *Comparison of the proposed algorithms.*

Method	MSE		# iterations	
	w/out NNLSE	w/ NNLSE	w/out NNLSE	w/ NNLSE
Fixed-Band	0.0008	0.0006	22867	25341
Fixed-σ	0.0038	0.0017	23801	26275
Fixed-μ	0.0054	0.0017	21841	24315
Fixed-$\{\mu, \sigma\}$	0.0098	0.0022	7434	22266

algorithm modifies the fundamental frequency f_0 of a harmonic sound while preserving its spectral envelope. This application is convenient to understand perceptually the spectral decomposition in Gaussian sums and is used to assess the accuracy of the model.

3.1 Harmonic plus Stochastic Model - HSM

In this work, the system of Erro et al. [9] is used as baseline. This one is a popular model for analysis, synthesis and modification of speech is known as *Harmonic plus Stochastic Model – HSM*. This robust and compact system for speech signal representation splits a signal into static frames and decomposes them in harmonic and stochastic parts.

If a frame s is evaluated as non-harmonic, that is, *unvoiced*, only the stochastic part s_e is returned and the frequency f_0 is set to zero. The spectral envelope is modeled from the smoothed version of the spectrum $|S_e|$, using cubic splines interpolation of spectral peaks. These peaks are estimated by applying of a maximum filter with a lag corresponding to $60Hz$ in the discrete spectrum. Then, the sum of Gaussians $G_e(a, \mu, \sigma)$ is obtained from $|S_e|$.

If s is evaluated as harmonic, both harmonic and stochastic parts s_h and s_a are returned. $G_e(a, \mu, \sigma)$ is estimated from the stochastic part by the same method used in the non-harmonic case. For the harmonic part, the method returns a set of L amplitudes A_l and corresponding initial phases ϕ_l for each harmonic $l \in \{1, \dots, L\}$, associated with a fundamental frequency f_0. We then produce a smoothed version of the harmonic spectrum $|S_h|$ by 15× upsampling the amplitudes A_l using cubic spline interpolation. This smoothed harmonic spectrum is then modelled using a sum of L Gaussians, obtaining a model $G_h(a, \mu, \sigma)$ that will be directly used for modifying the pitch.

Suppose that the desired pitch shifting factor is ρ, i.e., the new fundamental is $f_0' = \rho f_0$ and all harmonic components of the modified signal will have frequencies $f_l' = \rho f_l$, $\forall l$. In order to preserve the spectral envelope (and consequently the formants of the voice), we obtain the magnitudes A_l using downsampling of the spectrum $|S_h|$. The modulated reconstruction of all harmonic amplitudes from the sum of Gaussians $G_h(a, \mu, \sigma)$ is immediate: A_l' will be defined as $A_l' = [G_h(a, \mu, \sigma)] (f_l')$. Since the stochastic information is supposed to remain unchanged in this operation, it is sufficient to change the values A_l and ϕ_l of each frame to modulate the pitch.

Although the spectral envelope is supposed to remain unchanged, some care must be taken to ensure the naturalness of the processed sound, and the phase update is crucial in this process. Ideally, given the k-th frame it is desired to define $f_l^{(k)}$ so that $\phi_l^{(k)} = \phi_l^{(k-1)} + \frac{2\pi N}{R} f_l^{(k)}$. Moreover, phase fluctuations existing between adjacent frames must be preserved to maintain the naturalness of sound. Consider that

$$\hat{\phi_l}^{(k)} = \phi_l^{(k-1)} + \frac{2\pi N}{R} l f_0^{(k-1)}.$$

Defining the phase fluctuation as $d_{\phi_l}^{(k)} = \phi_l^{(k)} - \hat{\phi_l}^{(k)}$, we can rewrite $\phi_l^{(k)}$ as function of $d_{\phi_l}^{(k)}$, so that

$$\phi_l^{(k)} = \phi_l^{(k-1)} + (d_{\phi_l}^{(k)} + \frac{2\pi N}{R} l f_0^{(k-1)}) = \phi_l^{(k-1)} + \frac{2\pi N}{R} f_l^{(k)}. \tag{8}$$

It is seen that $f_l^{(k)} = l f_0^{(k-1)} + d_{\phi_l}^{(k)} \frac{R}{2\pi N}$.

For pitch shifting, $f_l^{(k)}$ is multiplied by a modulation factor, and the current initial phase is reconstructed from the previous one (Eq. 8) preserving the phase difference of the k-th harmonic, $d_{\phi}^{(k)}$.

4 Evaluation

A preliminary experiment was made using 8 sentences in Spanish, each recorded by 4 male and 4 female speakers. The experiment was split into two parts, corresponding to objective and subjective evaluations.

4.1 Objective Evaluation

In this preliminary experiment we have calculated Mean Square Error (MSE) between given synthetic signals with known spectral envelopes and estimated spectral envelopes using several envelope estimation techniques. The input signal used is a sawtooth signal, whose spectral envelope has a $\frac{1}{f}$ overall shape that can be directly obtained using the continuous-time Fourier Transform. Table 2 below presents MSE values and average number of iterations for each method. The threshold parameter used in these methods was $\epsilon = 10^{-5}$. In this experiment, the amplitude refinement step using NNLSE presented in Sec. 2.2 was not used, except in the Fixed-$[\mu, \sigma]$ technique, which is precisely NNLSE applied to the amplitudes of the Gaussian components. Besides the four novel methods presented in this paper, we also compare the results with the spectral envelopes obtained with Cepstrum and LPC models of similar size. As we can see, the Fixed-Band method presented the lowest errors, although Fixed-$[\mu, \sigma]$ may be more useful with respect to computational time.

This objective experiment considers the overall fitness of the spectral envelope for each proposed model. However, it might be useful to consider specifically the fitness of these models with respect to the harmonic peaks of a voiced signal. For this reason, a simple perceptual experiment has been made in order to subjectively evaluate the naturalness of a modulated signal.

Table 2. *Avg. MSE and Avg. # of Iterations for each method.*

Method	Avg. MSE	Avg. # of Iterations
Fixed-Band	0.0025	20693
Fixed-σ	0.0033	21496
Fixed-μ	0.0061	19753
Fixed-$[\mu, \sigma]$	0.0033	7340
CEPSTRUM(24)	0.0729	2384
LPC(24)	0.0318	2384

4.2 Subjective Evaluation

This subjective evaluation aims to analyse the naturalness of the modified voice signal after pitch modulation. The pitch shifting was chosen as an illustrative application due to the stronger link of pitch shifting with the spectral envelope.

The perceptual test consisted of pitch shifting 2 sentences uttered by 2 speakers by factors of 0.8 and 1.2. Table 3 presents the Mean Opinion Score (MOS) of 17 listeners for each method. Listeners were asked to rank the results of the different methods on a 5-point scale. The method contained in the baseline system of Erro et al. [9], using Discrete all-pole modeling (DAP) with 14 coefficients, that is a kind of LPC modeling using the autocorrelation of harmonic peaks; and the TD-PSOLA method, were also considered in this experiment.

Table 3. *MOS subjective evaluation.*

Method	Fem.(0.8)	Fem.(1.2)	Male(0.8)	Male(1.2)
Fixed-Band	4.44	3.69	2.94	3.69
Fixed-σ	3.94	3.88	2.75	3.69
Fixed-μ	3.69	3.38	2.56	3.25
Fixed-$[\mu, \sigma]$	3.94	3.81	2.56	3.75
DAP	3.40	3.20	3.81	3.88
TD-PSOLA	2.88	3.13	4.06	3.63

As can be seen in Table 3, these methods achieve better results with signals that contains fewer harmonics (female voices), but MOS values are lower on signals with many harmonics (male voices). These lower values may be due to problems in the phases estimation of each frame in the reconstruction step, and will be better addressed in future works.

5 Conclusion

In this paper[3], a new parametric spectral estimation method was introduced assuming that the spectral envelope can be represented as a sum of Gaussians.

[3] This research was sponsored by CAPES and FAPESP grant 2008/08632-8.

Four models were presented for modeling a continuous approximation of the spectral envelope, which is controlled by amplitude, center frequency and bandwidth parameters, which can be defined using fixed values or optimized as free parameters using a modification of Marquardt's algorithm. Two experiments were presented to evaluate these models within an objective framework using synthetic signals with ground-truth spectral envelopes, and in a subjective scenario considering a pitch shifting application of real voice signals. The results obtained indicate that sums of Gaussians seems to be a promising spectral modeling tool for representing spectral envelopes of voice signals.

References

1. P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of gaussians," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 1229–1232.
2. B. Nguyen and M. Akagi, "Spectral modification for voice gender conversion using temporal decomposition," *Journal of Signal Processing*, 2007.
3. E. Godoy, O. Rosec, and T. Chonovel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1313–1323, 2012.
4. H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2. IEEE, 1997, pp. 1303–1306.
5. P. Zolfaghari, S. Watanabe, A. Nakamura, and S. Katagiri, "Bayesian modelling of the speech spectrum using mixture of gaussians," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–553.
6. 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.
7. A. Goshtasby and W. 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. D. 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.
9. D. Erro, A. Moreno, and A. Bonafonte, "Flexible harmonic/stochastic speech synthesis," in *6th ISCA Workshop on Speech Synthesis*, 2007.
10. D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.