New Speech Enhancement Approach for Formant Evolution Detection

Jesus Bobadilla

Computer Science, U.P.M., Crta. de Valencia, Km 7, 28031 Madrid, Spain
jbobi@eui.upm.es

Abstract. Spectra is usually shown as a two-dimensional graph where colors are directly related to signal levels. A great deal of speech recognition work and research takes this type of parameter directly. In this paper we propose to combine typical signal level values with the vectorial components of a *Slope* matrix containing orientation information on spectra surfaces. This additional information will enable us to obtain an enhanced speech signal spectra as well as formant evolution detection and a matching method to compare speech spectra sections. The mathematical formalization is based on vector analysis and matrix operations, where the basic components are the normal vectors to a set of triangular surfaces covering the spectral values. This formalism enables the use of mathematical tools (Matlab or similar) in a very easy way; and from here it is possible to program algorithms and visualize the results efficiently.

1 Introduction

There are many techniques to enhance the spectra: S CHEUNG [1] proposes a combination of the wideband and narrowband spectra. Y. SHIN [2] suggests the use of spatial filters. V.R. CHARI [3] describes an adaptive method based on the slow change of the formants. K. KODERA [4] proposes the energy redistribution technique. Whilst D. KUNZ [5] describes a new spectral analysis transform, with results that are an improvement on Fourier's.

Speech sounds can be modeled as the vocal tract responds to a sequence of pulses. The resonance frequencies appear in the spectra with the greatest energy; these are the speech formants and their information is basic to spoken language recognition [6, 7, 8]. Formant detection provides useful information located between parameters and sounds, therefore it can be used to reduce the complexity of the necessary speech recognition Neural Networks (NN) or Hide Markov Models (HMMs). Formant detection facilitates the automatic parametric learning phase in speech recognition and makes speech modeling easier, providing a closer similarity to actual human speech.

In general, automatic speech recognition is based on parametric learning techniques, mainly HMMs or NN [9, 10, 11, 12]. The parameters used are usually LPC coefficients or FFT results [3, 6, 13, 14]. The quality of the results varies depending on the techniques applied and the aims desired (speaker dependent, speaker independent, large vocabulary, reduced vocabulary, isolated speech, continuous speech,

etc.). In all these situations we must face the conceptual gap that exists between the mathematical parameters and the human speech sounds.

2 Basics of the Method

Let's look at a time windowed Fourier analysis computed over a speech signal. The result can be shown as in Figure 1; where the *X*-axis represents time, the *Y*-axis represents frequencies and the *Z*-axis represents the spectral values obtained.

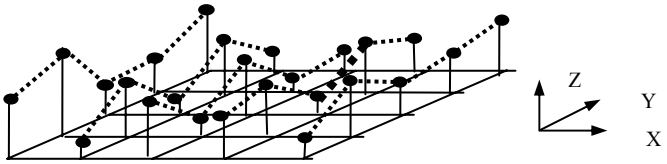

Fig. 1. Windowed Fourier analysis

In order to hold the spectral values (the *Z* magnitude at each point of figure 1) we will use a $F_{t,f}$ (Fourier $_{\text{time,frequency}}$) matrix:

$$F_{t,f} \quad f \in \left[0..\frac{N}{2}-1\right], \quad N = 2^n, \quad n \in \textit{Natural numbers} \tag{1}$$

In order to make a time and frequency evolution study of the spectral values contained in *F* we will consider only some of the original components of the matrix; then we will create a submatrix with *r* rows and *c* columns:

$$R_{r,c} \mid \; R_{r*\Delta r, c*\Delta c} = F_{t,f} \quad \Delta r, \Delta c \in \textit{Natural} \tag{2}$$

By varying the Δr and Δc parameters we can obtain different details in the time (*r*) and frequency (*c*) evolution estimations we are looking for. Figure 2 represents Figure 1 showing spectra using half of the values on the X-axis ($\Delta r = 2$) and the Y-axis ($\Delta c = 2$). The use of $\Delta r = \Delta c = 1$ has the effect of working with the whole $F_{t,f}$ matrix. In order to study time evolution speech characteristics (formants, etc.) we must increase the Δr parameter. In order to smooth frequency functions (as shown in Figures 1 and 2) we must increase the Δc parameter. The Δr and Δc parameters must not have high values; in this way we avoid excessive smoothing and the loss of determinant spectral signal peaks.

Fig. 2. Windowed Fourier analysis applying the Δr and Δc parameters

Using the desired Δr and Δc parameter values we obtain a set of points that can be studied as the basis of three-dimensional functions (Figure 3a). A different

approach would be to create a grid using triangles formed with the spectral value positions (Figure 3b). Triangles are the simplest geometrical shapes which fit the traditional spectral three-dimensional areas.

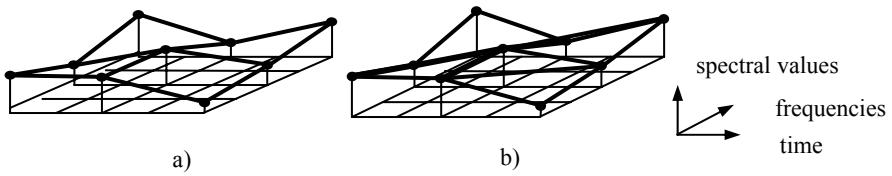

Fig. 3. a) Spectral areas of study. b) Areas of study as triangles

Now we have a reduced matrix $R_{r,c}$ containing the spectral values which will be the basis of a triangle-based envelope of the spectral information. These spectral values form a chosen subset of the Fourier $F_{t,f}$ matrix, designed to study the time/frequency evolution of spectral speech signals.

Starting from a matrix of points ($R_{r,c}$), it is possible to configure different dispositions of the triangle envelope. Figures 4a and 4b shows the most immediate and regular ones. Depending on the nature of the signal (speech, video frames, etc.) and the problem to solve, it can be more accurate to use different regular layouts as in Figure 4c or even irregular dispositions as showed in Figure 4d. We work with the regular layout presented in Figure 4b.

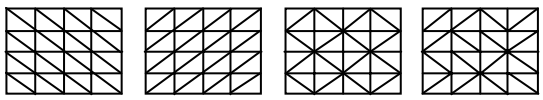

Fig. 4. Different dispositions of the triangle envelope

At this stage we have a triangle-based envelope of the spectra, designed to study time and frequency slopes. It is possible to characterize and study the time and frequency slopes of each triangle by comparing its three points, but it is simpler and more elegant to use the normal vector. Working with the normal vectors of the surfaces will allow us to compare the slopes of adjacent triangles, and, therefore, to calculate the time and frequency evolution of the speech signal.

To obtain the normal vectors we will use the following method:

1. We will use 4 auxiliary vectors:

$\vec{V}_{1x,y}\ \vec{V}_{2x,y}\ \vec{V}_{3x,y}\ \vec{V}_{4x,y}\quad \forall x \in 0..r-1,\ \forall y \in 0..c-1$, as shown in Figure 5

$$\vec{V}_{1\,x,y} = \left(1,0,r_{x+1,y} - r_{x,y}\right) \tag{3}$$

$$\vec{V}_{2\,x,y} = \left(0,1,r_{x,y+1} - r_{x,y}\right) \tag{4}$$

$$\vec{V}_{3\,x,y} = \left(1,0,r_{x+1,y+1} - r_{x,y+1}\right) \tag{5}$$

$$\vec{V}_{4\,x,y} = \left(0,1, r_{x+1,y+1} - r_{x+1,y}\right) \tag{6}$$

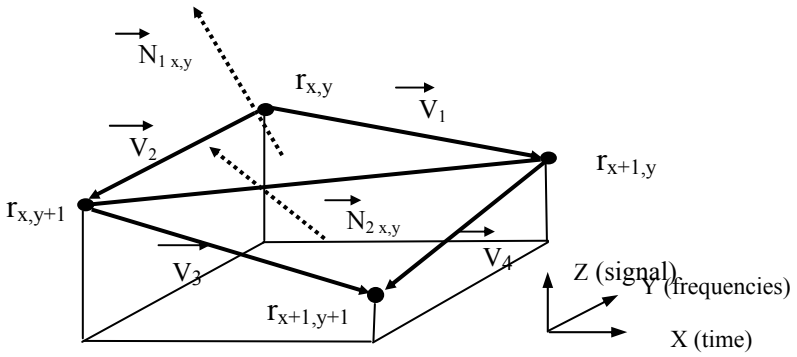

Fig. 5. Layout of the triangles and vector disposition

2. Using the auxiliary vectors we can compute the normal ones: $\vec{N}_{1x,y}$ $\vec{N}_{2x,y}$ $\forall x \in 0..r-1,\ \forall y \in 0..c-1$

$$\vec{N}_{1\,x,y} = \vec{V}_{1\,x,y} \otimes \vec{V}_{2\,x,y} \tag{7}$$

$$\vec{N}_{2\,x,y} = \vec{V}_{3\,x,y} \otimes \vec{V}_{4\,x,y} \tag{8}$$

In order to include absolute spectral speech signal information in the normal vectors, we will reflect the average spectral values of the three points of each triangle in its normal vector modulus:

$$\vec{N}_{1\,x,y} = \vec{N}_{1\,x,y} * \frac{r_{x,y} + r_{x,y+1} + r_{x+1,y}}{3} \tag{9}$$

$$\vec{N}_{2\,x,y} = \vec{N}_{2\,x,y} * \frac{r_{x+1,y+1} + r_{x,y+1} + r_{x+1,y}}{3} \tag{10}$$

This is the moment to create our final *Slope* matrix containing the normal vectors:

$$S_{2x,y} \quad x \in 0..2r-1,\ y \in 0..c-1 \text{ where } \vec{s}_{(2*x),y} = \vec{N}_{1\,x,y} \text{ and } \vec{s}_{(2*x)+1,y} = \vec{N}_{2\,x,y} \tag{11}$$

To get a better understanding of $S_{2x,y}$ matrix, we will use Figure 6. In Figure 6a we can observe a generic spectra fragmented using a triangle-based regular envelope. Each triangle has a normal vector associated; the modulus of these vectors are proportional to the average spectral signal value in the triangle areas. Figure 6b shows spectral time evolution using vectors; it is possible to get an analogous picture of the spectral frequency signal evolution; in this case we will select frequency-signal axes.

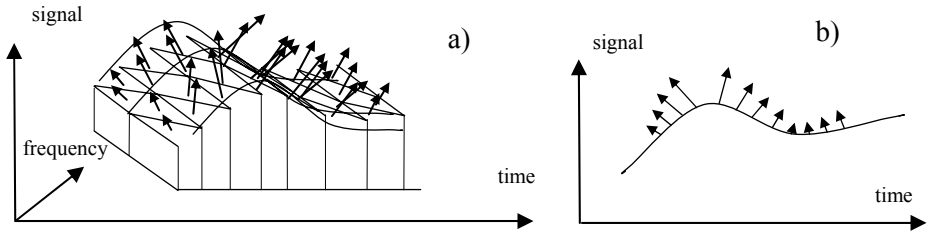

Fig. 6. a) $S_{2x,y}$ matrix b) Spectral time evolution

Figure 7 represents the matrix layout related to the triangle-based envelope. The link between matrix components and normal vectors has been established in (11).

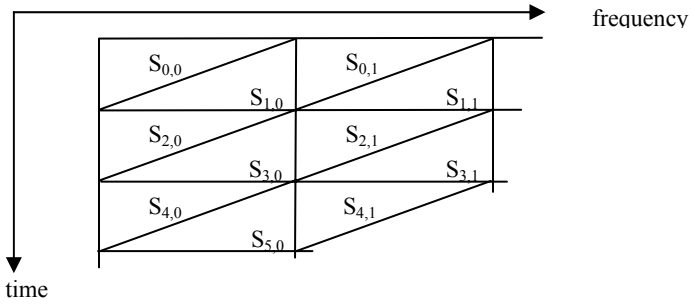

Fig. 7. Matrix layout related to the triangle-based envelope

3 Spectra Enhancement

Spectra is usually shown as a two-dimensional graph where colors are directly related to signal levels. A great deal of speech recognition work and research takes this type of parameter directly. In this paper we propose to combine the typical signal level values with the *Slope* matrix we created in the previous section. This additional information will allow us to obtain an enhanced speech signal spectra.

There are a wide range of studies and applications based on the correct determination of the speech formant position and evolution, perhaps the most important ones are linguistic studies, speaker detection, speech synthesis and speech processing in general. Using the typical methods, formant position and evolution determination are based on spectral peak detection; we will use the *Slope* matrix information for this purpose, looking for adjacent vectors forming an angle that is large enough to be considered to have been produced by a peak.

The most simple and mathematically elegant approach is to compute the inner product of all the adjacent components of the *Slope* matrix. Inside areas of maximums and minimums the inner product will be close to zero (perpendicular vectors on each side of the summit). With this idea in mind we can establish:

Formant detection 1:

$\forall x, \forall y$ (*except the border*) where we look for $\langle \vec{s}_{x,y}, \vec{s}_{x,y+1} \rangle \leq \theta$ (12)

In this way we not only detect maximums; but we also detect minimums. As we can easily obtain vector modulus, it is possible to enhance peaks and remove minimums using this information. Another way to achieve this goal is to compare frequency (Y-axis) angles:

Formant detection 2:

Let's use $\vec{s}_{x,y} = (x_1, y_1, z_1), \quad \vec{s}_{x,y+1} = (x_2, y_2, z_2)$

$\forall x, \forall y$ (*except the border*) where we look for:

$$\left(\arccos \frac{y_1}{\sqrt{x_1^2 + y_1^2 + z_1^2}} - \arccos \frac{y_2}{\sqrt{x_2^2 + y_2^2 + z_2^2}} \right) \geq \beta \qquad (13)$$

Formant evolution:

Formant evolution consists of the search of adjacent temporal peaks. As vocal tract has physic limitations, temporal peak evolution also has limitations. The translation of this fact to our model leads to the restricted search shown in Figure 8. For each triangle detected as a peak, we will search for temporal formant evolution on adjacent triangles located in an angle $90° \rangle \delta \rangle -90°$.

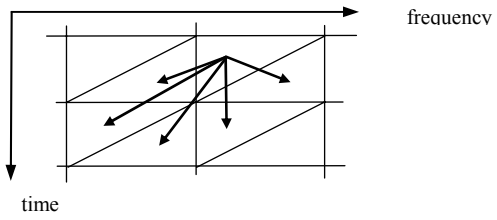

Fig. 8. **F**ormant evolution trajectories

$\forall \vec{s}_{x,y}$ detected as formant in (12) we can look for:

$$max \begin{Bmatrix} \langle \vec{s}_{x,y}, \vec{s}_{x+1,y} \rangle, \langle \vec{s}_{x,y}, \vec{s}_{x+2,y} \rangle, \langle \vec{s}_{x,y}, \vec{s}_{x+3,y-1} \rangle, \\ \langle \vec{s}_{x,y}, \vec{s}_{x+2,y-1} \rangle, \langle \vec{s}_{x,y}, \vec{s}_{x+1,y-1} \rangle \end{Bmatrix} \qquad (14)$$

By looking for the maximum value of these five inner products, we are searching for the adjacent triangle (Figure 8) that most closely resembles the studied triangle. This similarity is related to the slopes of the triangle (normal vector orientations), by looking for temporal continuity on the formant path.

4 Speech Matching

Speech recognition is usually based on a very large amount of speech samples that feed large neural networks or hidden Markov models. There are a variety of applications (such as computer-assisted second language learning, medical speech corrections, etc.) which do not need this heavy approach. In these cases, it is sometimes useful to have a light matching method to compare correct samples with the real time incoming ones. Based on the method presented in this paper, we will show a mathematical way to compute distances between pre-aligned and pre-normalized speech spectra sections. Using (11), we can establish a temporal (frequency limited) matching distance:

$$d_{t,f1,f2} = 1 - \left(\frac{1}{f2 - f1} \sum_{y=f1}^{f2} \left\langle \frac{\vec{s}_{t.y}}{\left|\vec{s}_{t.y}\right|}, \frac{\vec{s}_{t,y+1}}{\left|\vec{s}_{t,y+1}\right|} \right\rangle \right) \tag{15}$$

$$f2 > f1, \quad t \in 0..2r-1, \quad f1, f2 \in 0..c-1$$

Analogously, we can establish a frequency (time limited) matching distance:

$$d_{t1,t2,f} = 1 - \left(\frac{1}{t2 - t1} \sum_{x=t1}^{t2} \left\langle \frac{\vec{s}_{x.f}}{\left|\vec{s}_{x.f}\right|}, \frac{\vec{s}_{x+1,f}}{\left|\vec{s}_{x+1,f}\right|} \right\rangle \right) \tag{16}$$

$$t2 > t1, \quad t1, t2 \in 0..2r-1, \quad f \in 0..c-1$$

Finally we can establish a spectra rectangular section match:

$$d_{t1,t2,f1,f2} = \frac{1}{(t2 - t1)} \sum_{x=t1}^{t2} d_{t,f1,f2} = \frac{1}{(f2 - f1)} \sum_{y=f1}^{f2} d_{t1,t2,f} \tag{17}$$

$$d_{t1,t2,f1,f2} = 1 - \left(\frac{1}{(f2 - f1)(t2 - t1)} \sum_{y=f1}^{f2} \sum_{x=t1}^{t2} \left\langle \frac{\vec{s}_{x.y}}{\left|\vec{s}_{x.y}\right|}, \frac{\vec{s}_{x+1,y+1}}{\left|\vec{s}_{x+1,y+1}\right|} \right\rangle \right) \tag{18}$$

$$t2 > t1, f2 > f1 \quad t1, t2 \in 0..2r-1, \quad f1, f2 \in 0..c-1$$

The proposed matching equations are based on a simple slope-comparing method using inner products. In this case, d=1 means no matching at all. Perfect matching is d=0, and in real cases we look for $d \to 0$.

5 Conclusions

This paper shows a mathematical formalism based on vectorial notation; the formalism has been developed to facilitate the creation of new methods to achieve speech enhancement and formant evolution detection.

We use a *Slope* matrix containing orientation information on spectra surfaces condensed in vectorial notation. The *Slope* matrix contains enough data to be able to

work on the speech spectra, combining its individual elements and computing useful algorithms by using only simple vectorial notation.

The mathematical formalism presented in this paper enables the use of mathematical tools (Matlab or similar) in a really easy way; it is then possible to program algorithms and visualize results efficiently. This formalism facilitates new signal processing ideas that emerge in the speech research process.

Finally, our research group is now successfully using this formalism in its speech processing research; we are testing new original methods and implementing the algorithms with the vectorial facilities of Matlab. The result is a reduction in the time spent evaluating the new ideas and methods, as well as a general improvement in the research process.

6 References

1. Cheung S., Lim J. S.: Combined Multi-Resolution (Wideband/Narrowband) Spectrogram, ICASSP, (1991) 457-460
2. Shin Y., Choi H., Kim Ch.: A New Method For Enhanced Spectrogram of Speech, ICSP, 1997 (623-628).
3. Chari V. R., Espy-Wilson C. Y.: Adaptive Enhancement of Fourier Spectra, IEEE Trans. Speech and Audio Processing, vol. 3, (1995) 35-39
4. Kodera K., Gendrin R., Villey C.: Analysis of Time-Varying Signals with Small BT Values, IEEE Trans. Acoust., Speech, Signal Processing, vol. 26, (1978) 64-76.
5. Kunz D., Aach T.: Lapped directional transform: a new transform for spectral image analysis, Proc. ICASSP, (1999) 3433-3439
6. Rabiner L., Juang B. H.: Fundamentals of Speech Recognition, Prentice Hall, ISBN: 0130151572, (1993)
7. Bruce I.C., Karkhanis N.V.: Young E.D.: Robust Formant Tracking in Noise, Acoustics, Speech and Signal Processing (ICASSP), vol. 1, (2002) I13-I17
8. Yan Q., Vaseghi S.: Analysis, Modeling and Synthesis of Formants of British, American and Australian Accents, Acoustics, Speech and Signal Processing ICASSP, vol 1., (2003) 712-715
9. Lippman R. P.: Review of Neural Networks for Speech Recognition, Neural Computation , (1989), 1-46.
10. S. Simon, Neural Networks: A Comprehensive Foundation (2nd Edition), Prentice-Hall, 1998 ISBN: 0132733501
11. E. Varoglu, K. Hacioglu "Recurrent Neural Network Speech Predictor Based on Dynamical Systems Approach", Vision Image and Signal Processing IEE Proceedings, vol. 147, issue 2, pp 149-156, April 2000
12. S.M. Chu, T.S. Huang, "Audio-Visual Speech Modeling Using Coupled Hidden Markov Models", Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. IV4096-IV4099, May 2002
13. Gold B., Morgan N.: Speech and Audio Signal Processing, Wiley, (2000)
14. Gold B., Morgan N.: Speech and Audio Signal Processing; Processing and Perception of Speech and Music, John Wiley & Sons, ISBN: 0471351547, (1999)