

DETECTION OF NON-STATIONARITY IN SPEECH SIGNALS AND ITS APPLICATION TO TIME-SCALING

David Kapilow, Yannis Stylianou, Juergen Schroeter

AT&T Labs-Research, Shannon Laboratories, 180 Park Ave, Florham Park, NJ 07932-0971

{dak, yannis, jsh} @research.att.com

<http://www.research.att.com/projects/tts>

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ISCA Archive

<http://www.isca-speech.org/archive>

ABSTRACT

This paper describes an automatic method for the detection of non-stationarity in speech signals. It is based on three measures of non-stationarity using Line Spectrum Frequencies (LSFs), the derivative of RMS values, and a combination of these two features. The application of the proposed method to time-scaling of speech signals is also presented. Results from an informal listening test support its usefulness. Following these results, the method seems to be a powerful tool for the automatic control of time-scale factors based on the characteristics of the input speech signal. Listeners preferred our new method over applying a constant time-scale factor in 90% of all cases. Other possible applications of the proposed tool are also discussed.

1. INTRODUCTION

The detection of non-stationarity in speech signals has many important applications including speech synthesis, speech coding and speech recognition.

In speech synthesis, the majority of non-stationary portions of speech can be correlated with the phoneme boundaries. Hence, an estimation of non-stationarities may be used as an initialization step in phoneme labeling system. Another important application is the control of speech modifications (e.g., prosodic modifications, spectral smoothing) in current speech synthesis systems. In general, it is not desirable to modify or smooth highly non-stationary areas of speech; this introduces artifacts in the resulting synthetic speech signal. Finally, the measurement of non-stationarity can be used as an efficient way to evaluate concatenation algorithms objectively; in continuous speech and within a steady state voiced area, the proposed tool will measure a non-stationarity close to zero, while a high value of non-stationarity will be measured in concatenating two steady state voiced segments with different formant distributions left and right from the concatenation point. Therefore, an objective evaluation of smoothing algorithms has become possible (a high correlation between this objective way and a subjective listening test is very important).

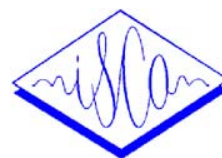
In speech coding, a measure of non-stationarity may be used in variable bit-rate speech coding. Time-adaptive analysis of speech is the base of the *temporal decomposition* method proposed by Atal [1]. A technique for localizing spectral stability events as an initialization to a temporal decomposition algorithm was recently presented

in [2].

For speech recognition, delta cepstrum and delta-delta cepstrum are used in an effort to capture dynamic characteristics of speech. This increases the dimension of feature vectors used for speech recognition considerably. However, using a simple scalar like the one proposed here, the non-stationarity characteristics of speech may be modeled without increasing the dimension of the input vector to the recognition system [3].

In [2], non-stationarities are detected by computing the transition rate of spectral parameters like Line Spectrum Frequencies (LSFs). While this method seems to work well for voiced sounds, it does not work well for detecting abrupt changes in the time domain (e.g., stop sounds). On the other hand, using the derivative of RMS values of speech signals, it is easy to detect this kind of changes. However, because the derivative of RMS values is a time domain criterion it does not detect variability in the frequency domain, such as the transition rate of certain spectral parameters. Indeed, the RMS based criterion is very noisy during voiced signals. Therefore, a combination of both criteria seems appropriate.

In this paper we propose a new criterion for the detection of non-stationarity in speech signals which combines the above two criteria. The proposed method is completely automatic and simple to apply. It has been evaluated in subjective tests by applying it to the automatic control of time scale modification of speech. The reason for our choice for this kind of evaluation was mainly based on the observation that a time expanded speech signal generated by a time domain technique (e.g., TD-PSOLA, WSOLA) suffers from unnatural artifacts and "tonalities". The amount of the perceived artifacts in a time scaled signal depends on the time scale factor (larger time scaling factors cause more occurrences of tonality) and on the speaker. On close inspection one finds that the areas most affected are those where the speech signal can be characterized as non-stationary. Therefore, if the time scale factors are under control on the areas of speech detected as non-stationary then one can expect to alleviate these tonalities from the time expanded signal. Results from an informal listening test support our hypothesis: in 90% of all cases, listeners preferred time-scaled files that were under the control of our stationarity measure over signals that were time-scaled with a constant time-scaling factor.



In the following sections, we first present the three criteria used for the detection of non-stationarity of speech signals. The three criteria are compared based on their performance to detect non-stationarity in labeled areas of speech signals. This is followed by the application of the criteria to the automatic control of time-scale factors during time-scaling of speech signals. Results from an informal listening test are presented in order to support our conclusion for the usefulness of the proposed criteria.

2. DETECTION OF NON-STATIONARITY OF SPEECH SIGNALS

2.1. First criterion, C^1

The first criterion is based on the transition rate of the RMS value, E_n , of the speech signal, $x[n]$, given by:

$$E_n = \sqrt{\frac{1}{N+1} \sum_{m=-N/2}^{N/2} x^2[n+m]} \quad (1)$$

A normalized transition rate of the RMS values is given by:

$$C_n^1 = \frac{|E_n - E_{n-1}|}{E_n + E_{n-1}} \quad (2)$$

which is the first proposed criterion. Based on this criterion,

$$C_n^1 = \begin{cases} \sim 1 & \text{if } |E_n - E_{n-1}| \text{ is large} \\ \sim 0 & \text{if } |E_n - E_{n-1}| \text{ is small} \end{cases} \quad (3)$$

Fig.1 shows two segments of speech and the corresponding transition rate of RMS values measured by Eq.2. From Fig.1(a) it is easy to see that this is a good criterion of detecting abrupt changes of speech signals (like the stop burst at time ~ 2.85 sec). However, we found C_n^1 is too sensitive in voiced speech¹. Indeed, Fig.1(b) shows how noisy this criterion is for voiced speech. Notice, that it doesn't detect the transition at time ~ 1.92 sec, which is a phoneme boundary.

2.2. Second criterion, C^2

This criterion is based on the gradient of the regression line for the evolution of Line Spectrum Frequencies (LSFs) over time. At instant n , within the time interval $[n \pm M]$, the gradient for the LSF, y_l , is given by:

$$g_n^l = \frac{\sum_{m=-M}^M m y_l(n+m)}{\sum_{m=-M}^M m^2} \quad (4)$$

for $l = 1, \dots, P$ where P is the number of estimated LSFs. Then, the measurement of transition rate is given by:

$$m_n = \sum_{l=1}^P (g_n^l)^2 \quad (5)$$

¹ Note that the window length for RMS computation is constant and independent of the characteristics of input speech signal, e.g., fundamental period. Maybe a pitch dependent window would result in a smoother derivative of RMS.

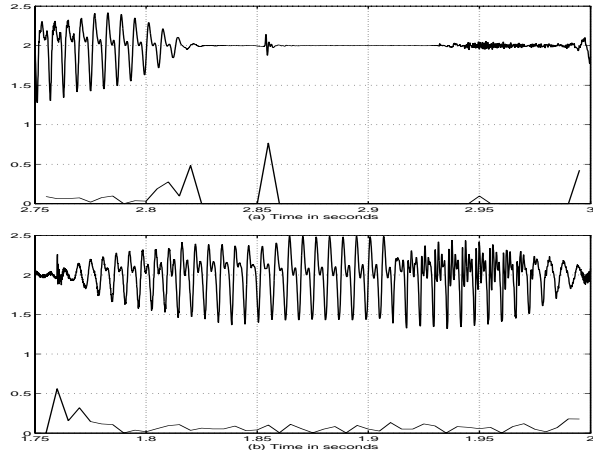


Figure 1: Results from the first criterion

Normalization of m_n between 0 and 1 is achieved through the function:

$$C_n^2 = \frac{2}{1 + e^{-\beta_1 m_n}} - 1 \quad (6)$$

In our experiments we used $P = 10$ and $\beta_1 = 20$. LSFs were calculated using a window of 30 ms. at 5 ms. frame rate, using a standard Linear Prediction procedure (a pre-emphasis of the signal prior to the computation of the AR filter has been used). The weight β_1 was defined based on measurements made on known stable and voiced signals and on signals with fast transitions.

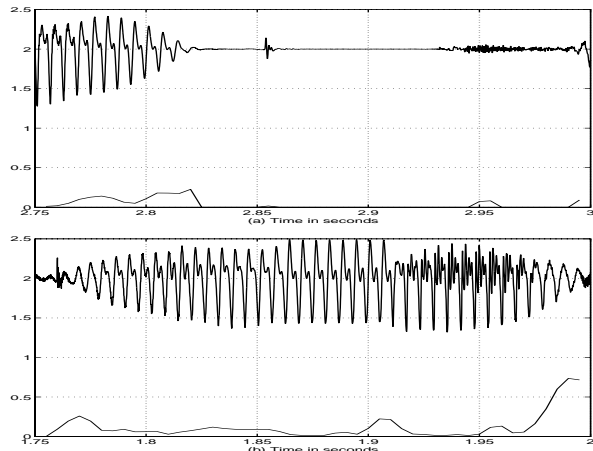


Figure 2: Results from the second criterion

Fig.2 shows the same segments of speech as Fig.1 along with C_n^2 . In Fig.2(top) the stop signal is not detected while in Fig.2(bottom) the transition rate of spectral parameters is smoother than the transition rate of the RMS values from criterion C_n^1 . Moreover, this criterion seems to be more discriminative in voiced sections (e.g., ~ 1.92 sec) than the first criterion. However, while this criterion is useful for voiced sounds it is not appropriate for sounds like stops (or, in general, speech events with short duration) because the gradient of the regression line in these cases is close to zero. Thus, we found that a combination of criteria C_n^1 and C_n^2 is more appropriate for the detection of all kinds of transitions in speech.

2.3. Third criterion, C_n^3

In an attempt to improve the performance of the first and second criteria, a third criterion is proposed which is the combination of the two previous criteria:

$$C_n^3 = \frac{2}{1 + e^{-\beta_2 m_n - \alpha C_n^1}} - 1 \quad (7)$$

where $\beta_2 = 17$ and α is given by:

$$\alpha = \begin{cases} 18.43(1.001 - 1.0049 e^{C_n^1} + C_n^1 e^{C_n^1}) & \text{if } C_n^1 \leq 0.5 \\ 0.5 & \text{if } C_n^1 > 0.5 \end{cases} \quad (8)$$

The values of parameter α in Eq. 8 have been determined by a least squares approach trying (using an exponential model) to normalize the criterion between 0 and 1. Fig.3 shows the same speech segments as Fig.1 (and Fig.2) along with the results from the third criterion. It is easily seen that criterion C^3 combines C^1 and C^2 .

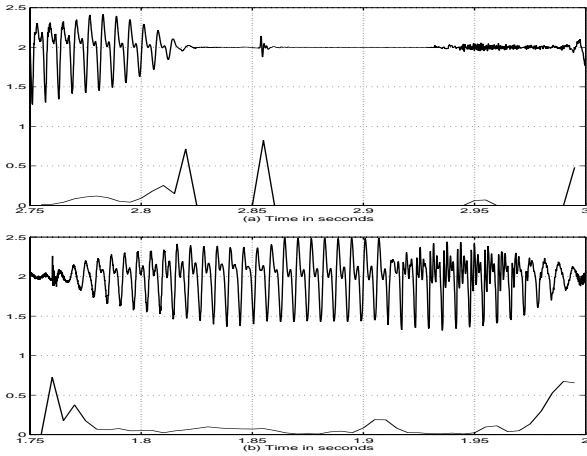


Figure 3: Results from the third criterion

2.4. Example

The performance of the three criteria has been tested on a big speech database including male and female speakers. In all these speech signals the behavior of C^1 and C^2 was similar to the one described in the previous sub-sections, while the third criterion has been found to successfully combine the criteria, C^1 and C^2 . An example of applying the three criteria to a speech signal uttered by a female speaker is shown in Fig. 4.

3. APPLYING THE PROPOSED CRITERIA TO THE TIME-SCALING OF SPEECH SIGNALS

Simple and flexible time domain techniques have been proposed (e.g., TD-PSOLA [4], WSOLA [5]) for time scaling of speech signals. While the quality of the time-scaled signal is good for time-scaling factors close to one, a degradation of the signal is perceived when larger modification factors are required. The degradation is mostly perceived as tonalities and artifacts in the stretched signal. These

tonalities do not occur everywhere in the signal, but are localized in transitional segments. Therefore, if there is a function $f(t)$ with the following characteristics

$$f(t) = \begin{cases} \sim 0 & \text{when a speech segment is stationary} \\ \sim 1 & \text{when a speech segment is non-stationary} \end{cases} \quad (9)$$

then a time scale factor β can be controlled by $f(t)$ in the following way:

$$\beta = 1 + d(t)b \quad (10)$$

where $d(t) = 1 - f(t)$ and b is the desired relative modification of the original duration. For example, if a signal is stretched by 25%, $b = 0.25$ and β without any control is $\beta = 1.25$. If the speech segment under stretching is stationary then $d \simeq 1$ and $\beta \simeq 1+b$. In case that the segment is non-stationary then $d \simeq 0$ and $\beta = 1$ which means that no modifications are permitted to this speech segment. Fig.5 shows an ideal function $f(t)$ with the above characteristics. Because all of the above criteria (C^1 , C^2 , and

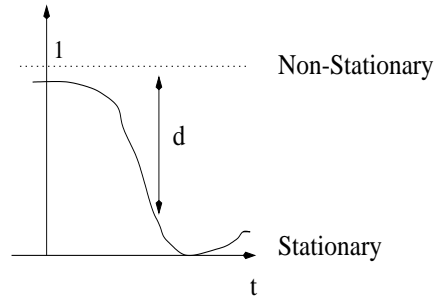


Figure 5: Ideal function $f(t)$.

C^3) have characteristics similar to this desired function, it was decided to apply these criteria to control time scaling.

For the purpose of this test a simple and efficient time-domain approach for time scaling of speech signals proposed in[5] was used. The corpus of the test included three male voices and three female voices recorded at a sampling frequency of 16kHz. Listeners were free to choose either headphones or speakers, and listen in their own environments (offices). The listening test was run through a web site and listeners were allowed to hear the files as many times as they wanted. 23 listeners participated. Not all of the listeners had experience listening to time-scaled signals. In order to simplify the listening test only criteria C^1 and C^3 were considered. Table.1 shows the time scale factor applied to six speech files and the preference of the listeners. Overall, in 90% of the selections listeners preferred the time-scaled sentences that were under control. Note that in the remaining 10%, most of the “no-control” preferences came from the time-compressed sentence ($\beta = 0.5$) where it was very difficult to detect any differences. However, in case of time expansion the preference for the controlled time scaled signals is clear. Considering only the results from the time expanded signals there is a 94% preference for controlling time-scaling with C^1 or C^3 . There is a strong positive correlation between modification factor, β , and preference of listeners for criterion C^3 ($r = 0.68$) and for both criteria C^1 and C^3 together ($r = 0.89$).

The quality of time-expanded signals was judged to be

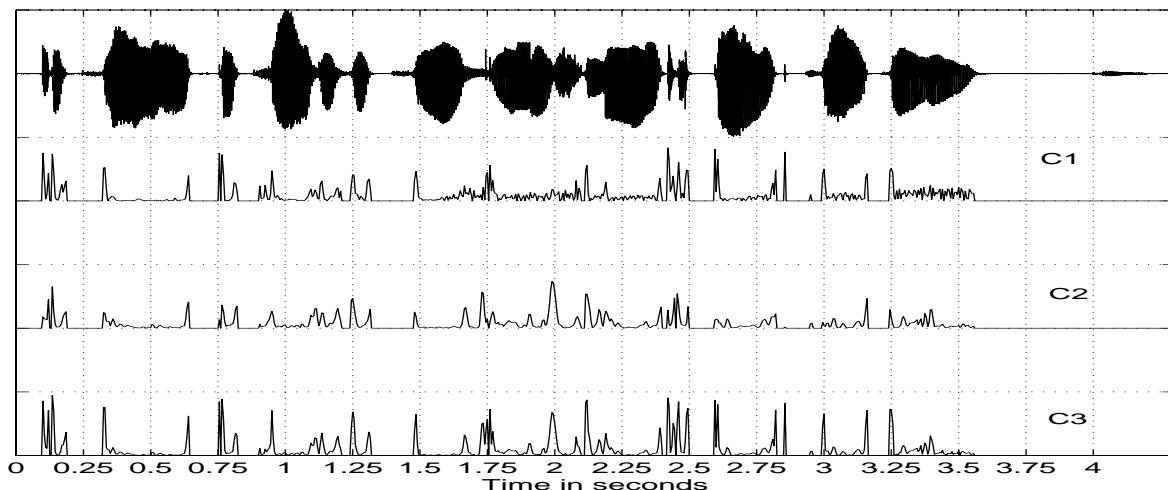


Figure 4: Measurement of Transition Rate in Speech using the three criteria

	File 1	File 2	File 3	File 4	File 5	File 6
Time scale factor	$\beta = 0.5$	$\beta = 2$	$\beta = 2$	$\beta = 4$	$\beta = 4$	$\beta = 4$
No control	30%	17%	9%	4%	0%	0%
Control with C^1	26%	57%	65%	39%	70%	22%
Control with C^3	44%	26%	26%	57%	30%	78%

Table 1: Results from the formal listening test of controlled time-scaling using criteria C^1 and C^3

almost free of any artifact (e.g., tonality). While C^1 and C^3 had similar average scores (46.5% and 43.5%, respectively), we believe that criterion C^3 is more general than C^1 . Moreover, criterion C^3 is more suitable for other applications already mentioned in the Introduction, while the usage of criterion C^1 is rather limited. For instance, in contrast with C^1 , criterion C^3 seems to indicate phoneme boundaries.

4. CONCLUSION

In this paper we have presented a method to automatically detect non-stationarities in speech signals. The method combines the transition rate of the RMS values (time-domain criterion) and the transition rate of spectral information (frequency domain criterion). The proposed method was subjectively evaluated by its application to control time scale factors in order to remove tonalities and artifacts in the time expansion of speech signals performed by simple time-domain techniques (e.g., WSOLA, TD-PSOLA). An informal listening test has shown that listeners prefer the control of time scale factors in areas where non-stationarity was detected using the proposed tool. Based on the fact that, in a time-expanded signal, tonalities are perceived in non-stationary areas of speech, and that these artifacts were removed when the proposed method was applied, we may conclude that our method successfully alleviates this problem.

The application of the proposed criterion in other areas of

speech processing is under investigation. In speech synthesis, the criterion may be useful in deciding whether to apply spectral smoothing.

5. REFERENCES

- [1] B. Atal, "Efficient coding of lpc parameters by temporal decomposition.," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 81–84, 1983.
- [2] A. Nandasena and M. Akagi, "Spectral stability based event localizing temporal decomposition.," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, (Seattle, USA), pp. 957–960, 1998.
- [3] S. Furui, "On the role of spectral transition for speech recognition," *J. Acoust. Soc. Am.*, vol. 80, no. 4, pp. 1016–1025, 1986.
- [4] E. Moulines and W. Verhelst, "Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), pp. 519–555, Elsevier, 1995.
- [5] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," *Proc. IEEE ICASSP-93*, pp. 554–557, 1993.