Correlation of STFT Spectrograms Applied to the Voice Deal Function in Mobile Phones

JOÃO S. PEREIRA[1], RUI V. B. MONTEIRO[1]
JOSÉ MAGNO LOPES[1] e HENRIQUE J. A. DA SILVA[2]

[1] Departamento de Engenharia Informática,
Escola Superior de Tecnologia e Gestão, Alto do Vieiro,
Morro do Lena, P-2401-951 Leiria,
Apartado 3063,
PORTUGAL

[2] Instituto de Telecomunicações,
Universidade de Coimbra,
Polo II, P-3030-290 Coimbra,
PORTUGAL

Abstract: - For the past five decades there has been a substantial increase in the development of computer applications based on the STFT (Short-Time Fourier Transform) spectrograms. Amongst others, a relevant example of this kind of implementation is speech analysis. Indeed, if the audio signal is correctly filtered, it is possible to use STFT spectrogram techniques to recognize speech. This work investigates the use of STFT spectrograms in speech recognition, using the voice deal function of mobile phones. The voice recognition rate obtained is then compared with that obtained with some commercially available mobile phones. The high recognition rate obtained validates the use of a STFT spectrogram methodology in speech recognition and certainly encourages further work to be carried out.

Key-Words: STFT, Spectrogram, Voice Recognition, Correlation.

1. INTRODUCTION

The most straightforward approach to characterize the frequency of a signal as a function of time is to divide the signal into several blocks that may be overlapped. Then, the Fourier Transform can be applied to each block of data in order to identify its frequency contents. This process has become known as the short-time Fourier Transform (STFT) [1,2] and roughly reflects how the signal frequency contents changes over time. The size of the blocks determines the time accuracy: the smaller the block, the better the time resolution. However, frequency resolution is inversely proportional to the size of a block. While small blocks yield good time resolutions, this deteriorates the frequency resolution and vice-versa. Traditionally, this phenomenon is known as the window effect.

The STFT spectrogram is one of the quadratic JTFA (Joint Time-Frequency Analysis) methods [3] and, as previously mentioned, the STFT spectrogram suffers from the window effect. Several algorithms have been developed to minimize this problem, such as:

- Adaptive spectrogram,
- Cohen's class,
- Choi-Williams distribution,
- Cone-shaped distribution,
- Wigner-Ville distribution,
- Gabor spectrogram.

By employing one of these methods, it is possible to process signals that conventional Fourier Transform cannot handle.

A periodic signal can be expressed by means of its Fourier Series. The Gabor expansion represents a signal *s[i]* as the weighted sum of the frequency-modulated and time-shifted function *h[i]*:

$$s[i] = \sum_{m} \sum_{n=0}^{N-1} C_{m,n} h[i - m\Delta M] e^{\frac{j2\pi i}{N}} \tag{1}$$

where the Gabor coefficients $C_{m,n}$ are computed by the STFT [4]:

$$C_{m,n} = STFT[m\Delta M, n] = \sum_{i=0} s[i]\gamma^*[i - m\Delta M]\, e^{\frac{-j2\pi i}{N}} \qquad (2)$$

The STFT-based spectrogram is defined as the square of the STFT:

$$SP[m\Delta M, n] = \left| \sum_{i=0} s[i]\gamma[i - m\Delta M] e^{-\frac{j2\pi ni}{N}} \right|^2 \qquad (3)$$

The γ function represents the Hanning Window:

$$\gamma_i = 0.5x_i[1 - \cos(\omega)] \qquad (4)$$

$$\omega = \frac{2\pi i}{N}, \quad i = 0, 1, 2, ..., \mathrm{N - 1} \qquad (5)$$

where N denotes the number of frequency bins, and ΔM denotes the time sampling interval. The STFT-based spectrogram is simple and fast, but suffers from the window effect. This problem can be minimized if a Hanning Window previously filters each data block.

2. WORD RECOGNITION

2.1 Word spectrogram

We have applied the STFT to audio signals by means of the correlation function, for speech recognition. The algorithm implemented includes the following steps:

1- Voice signal acquisition: signal sampled at 22050Hz with 8 bits and no DC component.

2- Signal smoothing for reducing undesirable high frequencies: each signal point becomes the average of the 5 consecutive preceding points.

3- Detection of the beginning and end of each word with the remove of silence zones.

4- New transformation: each point is made equal to the difference between the actual point and the second preceding point.

5- Further smoothing: each signal point becomes the average of the 3 consecutive preceding points.

6- Signal clipping at 30% of its maximum value.

7- For an audible signal, the maximum value is made equal to 127.

8- The signal is divided into several blocks with 256 points. Each block is overlapped by 128 points of contiguous blocks.

9- Each block is filtered by a Wideband Hanning Window.

10- Calculation of the square of the Fourier Transform in all blocks.

11- Amplification of high frequencies so that the amplitudes of all frequencies are equal. The filtering function is equal to $f^{7/2}$.

12- Remove less audible frequencies: f<f(0)=86Hz and f > f(35)=3100Hz.

13- Application of a base 10 logarithmic function with clipping around the average value.

14- STFT spectrogram smoothing: each spectrogram point is equal to the average of the actual point and the 8 points around it. This smoothing operation is repeated 4 times.

15- Independently of the time size of each word, all distinct spectrograms will have 128 blocks. This time normalisation is achieved by means of an interpolation function.

16- Calculation of partial derivatives in both time and frequency domains of the normalised STFT spectrograms.

17- Partial derivative smoothing: each spectrogram point is equal to the average of the actual point and the 8 points around it.

18- Amplitude normalisation of all these new blocks.

19- Sum of these two new partial derivative spectrograms.

20- Spectrogram translation of all points to the region around the average value, with amplitude normalisation.

Note: theoretical details are not included here due to lack of space, but may be provided on request.

All these steps originate a special spectrogram, which is illustrated in Figure 1. This plot illustrates the spectrogram of the word “John”. The Figure 2 is a 2nd spectrogram of the same word “John”. As we can see there is a similitude between the two spectrograms.

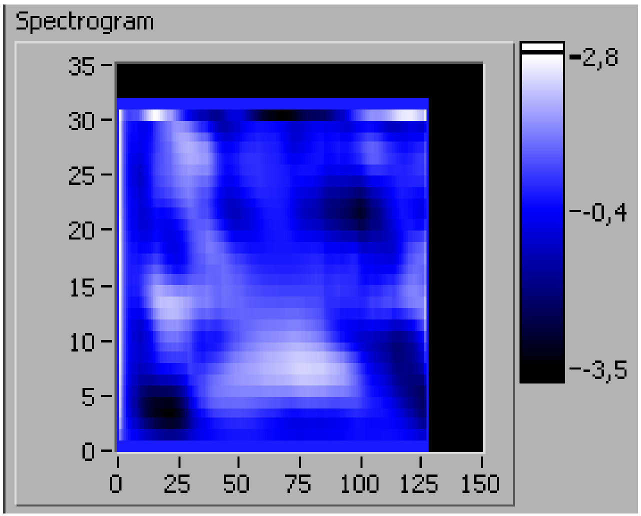


Figure 1: Final spectrogram of the word "John" (1st acquisition).

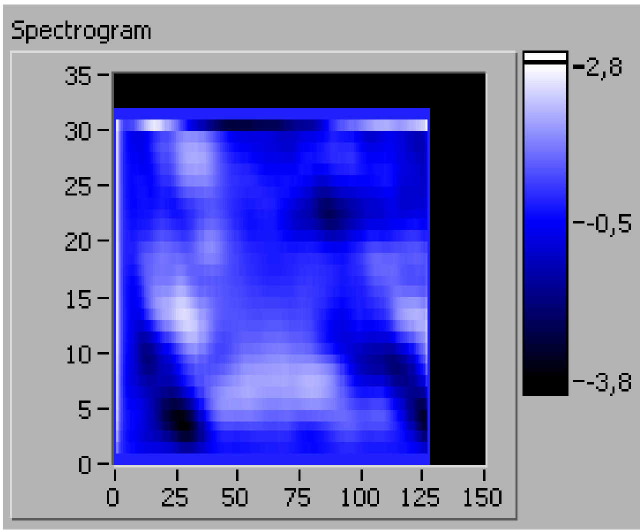
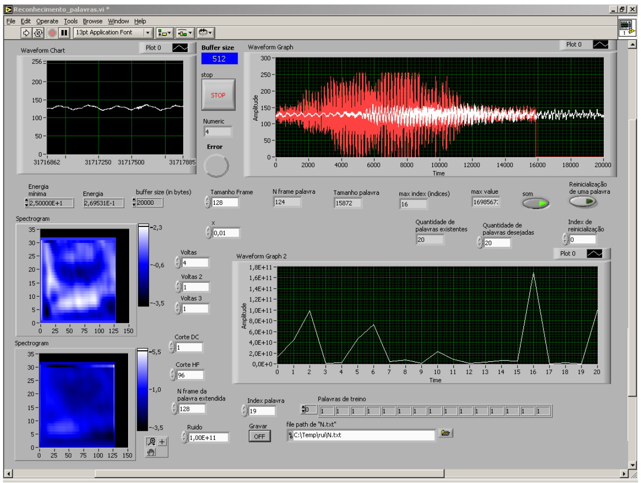


Figure 2: Final spectrogram of the word "John" (2nd acquisition).

Speech recognition itself is achieved by a partial correlation function amongst STFT spectrograms. It is necessary to create a database with referential STFT spectrograms that includes all words in order to match the correct spectrogram. The database selected contained 20 names of a mobile phone agenda. The maximum value of partial correlation indicates the higher similarity between STFT spectrograms, meaning that the right word has been recognised.

The STFT spectrogram is separated in 4 parts (32 blocks) in time order. For example: the STFT spectrogram of the word "ABCD" is divided in 4 distinct spectrograms ("A", "B", "C", and "D"). The product of the 4 partial correlations (of the 4 distinct partial spectrograms) is then the total correlation value.

The speech recognition application was developed using National Instruments' LabViewTM version 6.0. Its front panel is illustrated in Figure 3.



Buffer size 512

Numeric 4

Energia mínima 2,50000E+1

Energia 2,69531E-1

buffer size (in bytes) 20000

Tamanho Frame 128

N frame palavra 124

Tamanho palavra 15872

max index (indices) 16

max value 1698567:

x 0,01

Quantidade de palavras existentes 20

Quantidade de palavras desejadas 20

Index de reinicialização 0

Voltas 4

Voltas 2 1

Voltas 3 1

Corte DC 1

Corte HF 96

N frame da palavra extendida 128

Ruido 1,00E+11

Index palavra 19

Gravar OFF

Palavras de treino 0 | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

file path de "N.txt" C:\Temp\rui\N.txt

Figure 3: Speech recognition application front panel.

2.2 Speech Recognition Rate

This STFT spectrogram methodology may be used in mobile phones with a voice deal function. The database selected contained 20 random Portuguese names of a mobile phone agenda. The STFT spectrogram calculation was obtained using these 20 names. Three different users (2 men and 1 woman) have tested this algorithm, as this methodology is supposed to be user dependent. Each user has separately tested 20 recognition names several times, and the following results were obtained:

STFT spectrogram correlation:
1st user = 94,5%
2nd user = 95,0%
3rd user = 92,0%
Average = 93,8%

Mobile phone – model S:
1st user = 87,5%
2nd user = 89,0%
3rd user = 90,0%
Average = 88,8%

Mobile phone – model M:
1st user = 57,0%
2nd user = 70,0%
3rd user = 78,0%
Average = 68,3%

3. CONCLUSION

Nowadays, many applications make use of automatic speech recognition, and a practical example is the voice deal function of mobile phones. However, this function is usually characterized by a limited efficiency. In order to overcome this limitation and improve the recognition efficiency, several recognition methods have been proposed. This work focuses on one of these methods, entitled "Correlation of STFT Spectrograms", and its possible application to the voice deal of a mobile phone.

The studies carried out and presented here demonstrate that the speech recognition rate obtained with the STFT Spectrogram Correlation is more successful than the algorithms implemented in commercial mobile phones. If this method could be implemented in this kind of devices, it should be possible to increase the mean value of the speech recognition rate at least 5%. This figure was experimentally achieved with 20 distinct Portuguese names, where each word spectrogram is the mean of two spectrograms of the same word.

The high recognition rate obtained validates the use of the STFT spectrogram methodology in speech recognition and certainly encourages further work to be carried out. The Hidden Markov Models (HMM) are an example of the application of STFT Spectrograms. This methodology anticipates optimistic results in isolated word recognition based on previous experiences with Linear Predictive Coding (LPC) and Linear Predictive Ceptrum Coefficient (LPCC), when HMM [5] are applied.

4. REFERENCES

[1] "Signal processing Toolset", part I – Joint Time–frequency Analysis Toolkit, *National Instruments*, January 1999.

[2] Choi, H., and W. J. Willians, "Improved time-frequency representation of multicomponents signals using exponential kernels" *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37.6 (1989): 862-871.

[3] Qian, S., and D. Chen, "Joint time-frequency analysis" Prentice-Hall, Englewood Cliffs, N.J., 1996.

[4] Wexler, J., and S. Raz, "Discrete Gabor expansions", *Signal Processing*, vol. 21.3 (1990):207-221.

[5] João S. Pereira[1], Alexandre J. M. Santos, and Fernando M. S. Perdigão, "Real Time Recognition of Isolated Words Using Non Observable Markov Models", final year project dissertation, Department of Electrical and Computer Engineering, Faculty of Sciences and technology, University of Coimbra, Portugal, 1992 (in Portuguese).