Double Fourier analysis for Emotion Identification in Voiced Speech

D. Sierra-Sosa[1], M. Bastidas, D. Ortiz P., and O.L. Quintero

Mathematical Modeling Research Group, GRIMMAT, School of Sciences, Universidad EAFIT, Carrera 49 N° 7 Sur-50, Medellín – Colombia.

E-mail: dsierras@ eafit.edu.co

Abstract. We propose a novel analysis alternative, based on two Fourier Transforms for emotion recognition from speech. Fourier analysis allows for display and synthesizes different signals, in terms of power spectral density distributions. A spectrogram of the voice signal is obtained performing a short time Fourier Transform with Gaussian windows, this spectrogram portraits frequency related features, such as vocal tract resonances and quasi-periodic excitations during voiced sounds. Emotions induce such characteristics in speech, which become apparent in spectrogram time-frequency distributions. Later, the signal time-frequency representation from spectrogram is considered an image, and processed through a 2-dimensional Fourier Transform in order to perform the spatial Fourier analysis from it. Finally features related with emotions in voiced speech are extracted and presented.

1. Introduction

Spectrograms have being widely used over the years due to the broad band of applications in several areas of signal processing [1-5]. In voice speech signals, spectrograms allows for identify subtle cues related with utterance acoustics, providing the means to identify the speaker, separate speech from background noises, recognize and transcript spoken word to text, among others [6-7]. A particular characteristic when synthesizing speech recordings by using spectrograms, is formant transitions occurrence, this structures often related to consonants in spoken word, arrange in well-defined patterns and groups portraying speech features, which can be analyzed to extract information from signal [1, 6, 8].

Image processing techniques can be employed to analyze spectrograms, given the particular structures from power spectrum densities distributions [9-14]. By considering spectrograms as images, several features from signals can be extracted and recognized. For instance, the principal components can be classified in terms of Fourier descriptors with a high recognition rate [7]. Nonetheless, Fourier analysis provide either good temporal accuracy or good frequency accuracy, leading to other approaches as multi-scaled filtering and correlation methods to synthesize variations in color, texture and shape [15]. Also, mathematics morphology operators like erosion and dilation have being used to reduce spectrogram noise, which combined with two-dimensional non-linear filtering, allows for automatic speech recognition and enhancement [10, 11].

Signal processing derived from the time-frequency representation of spectrogram images, by following image processing methods, allows for identify and recognize features induced by emotions in voiced speech [16, 17]. We propose an analysis technique to identify emotion related information

on spectrograms, based on the spatial Fourier analysis from binary threshold spectrogram images preserving the formant transition structures. Double Fourier analysis results, related with the particular characteristic from each synthetic emotion in Berlin Database of Emotional Speech are presented. This database includes ten different phrases, spoken by ten different speakers, portraying six synthetic emotions: Anxiety, Disgust, Happiness, Boredom, Anger and Sadness.

2. Proposed Method

From recorded voiced speech signals, spectrograms are obtained by using short time Fourier Transform (STFT), with different Gaussian windows, depending on signal sampling. These windows allows for obtain sufficient accuracy in both time and frequency, which resolution limits are described by Heisenberg's uncertainty principle [18]. Also, the windows are overlapped a 50% in order to adequately describe signal changes.

The spectrogram power spectral density information is converted to a gray-level image, where each matrix position is considered a pixel with values ranging from 0 to 255 [19]. The latter is then converted to a 0 and 1 valued matrix by using a binary threshold [20]. This procedure allows for preserve the spatial distribution and shape information from formants. Despite the evident information loss in the binary threshold process, the formant shape and spatial distribution is sufficient information to relate speech acoustic cues with emotion induce features in voiced speech.

A spatial 2D Fourier Transform is obtained from the binary image processed from each voiced speech record. The power spectral densities from these spatial transforms portray particular features, related with each synthetic emotion available in Berlin Database of Emotional Speech. These power spectral functions are considered as new gray-level images, containing the spatial frequency distributions from each spectrogram image. An average image is obtained for each emotion. In these average images the emotion-related feature become apparent.

Figure 1 depicts the proposed signal processing method. From voiced speech audio recording a STFT is obtained, the signal is considered as an image in following steps and synthesized accordingly. The power spectral density gray level image is binarized using a threshold, to extract formant shape and distribution information. A spatial 2D Fourier transform is obtained by means of a Fast Fourier Transform algorithm, and an average image from power spectra function is obtained for each synthetic emotion. A vertical line profile, crossing the average image origin, is evaluated. A qualitative evaluation from line profiles allows for infer that emotion-induced features in voiced speech may be extracted when synthesized by the proposed method; as each emotion in Berlin Database of Emotional speech contain recordings from both different speakers and phrases.

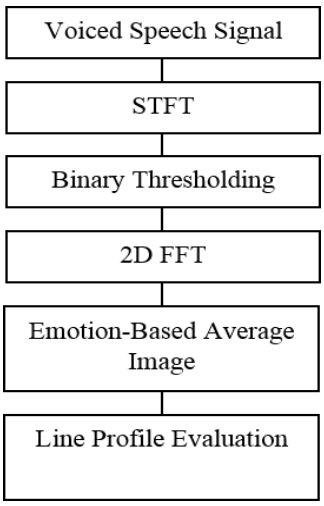


Figure 1. Block Diagram for voiced speech synthesis and analysis using double Fourier method

3. Results

The results obtained by the Double Fourier analysis are depicted in Figure 2, where the rows in descending order display the spatial power spectrum densities corresponding with the synthetic emotions anxiety, disgust, happiness, boredom, anger and sadness respectively; the columns are related with audio sampling and STFT windows parameters, in the first column the results obtained with voiced speech recordings sampled at 8 KHz with a 256 samples sized Gaussian window are presented, in the second and third column portray the results for recordings sampled at 16 KHz, with 512 and 1024 samples sized Gaussian windows respectively. All Gaussian window where overlapped a 50%.

Albeit slight changes in the spatial spectral density distributions, induced by voiced speech sampling and STFT window parameters. Each synthetic emotion portrays a recognizable particular structure.

When the recordings are processed by using the proposed Double Fourier Analysis, the emotion particular structures become apparent. These shapes, when conducting a qualitative analysis, are independent from audio sampling and STFT Gaussian window parameters. The latter, broaden the spatial frequencies distribution, but each particular structure is preserved.

The spatial spectral density distribution qualitative analysis was conducted by synthesizing the vertical line profiles from averaged images. In these line profiles, particular structures become apparent for each synthetic emotion, which cannot be related with speakers or phrases, because both different speakers and phrases were included in the Double Fourier Analysis process. Figure 3 present the vertical line profiles for each emotion, the left column depict the results related with emotional voiced speech sampled at 8 KHz, with 256 samples sized Gaussian window in the STFT; the right column present the results when using recordings sampled at 16 KHz, with 512 samples sized Gaussian window. Both windows were overlapped a 50%.

From figure 3, it should be noted that, despite slightly changes in the line profiles corresponding to each sampling rate, both results portray characteristics that can be tracked over particular emotional speech recordings, such as local maxima and local maxima positions, center peak width and shape. With this information a particular line profile shape can be related with each emotion in Berlin Database of Emotional Speech.

An important fact is that those particular profiles associated to each emotion (extracted through our double Fourier approach) can be easily introduced to a classifier in order to perform the task of emotion recognition.

We hypothesize that this methodology can be used as emotion recognizer independent from language because of the physical features extracted from the signal in both time frequency domain and spatial frequencies.

Figure 2. Spatial power spectral densities obtained by the Double Fourier Analysis of emotional voiced speech

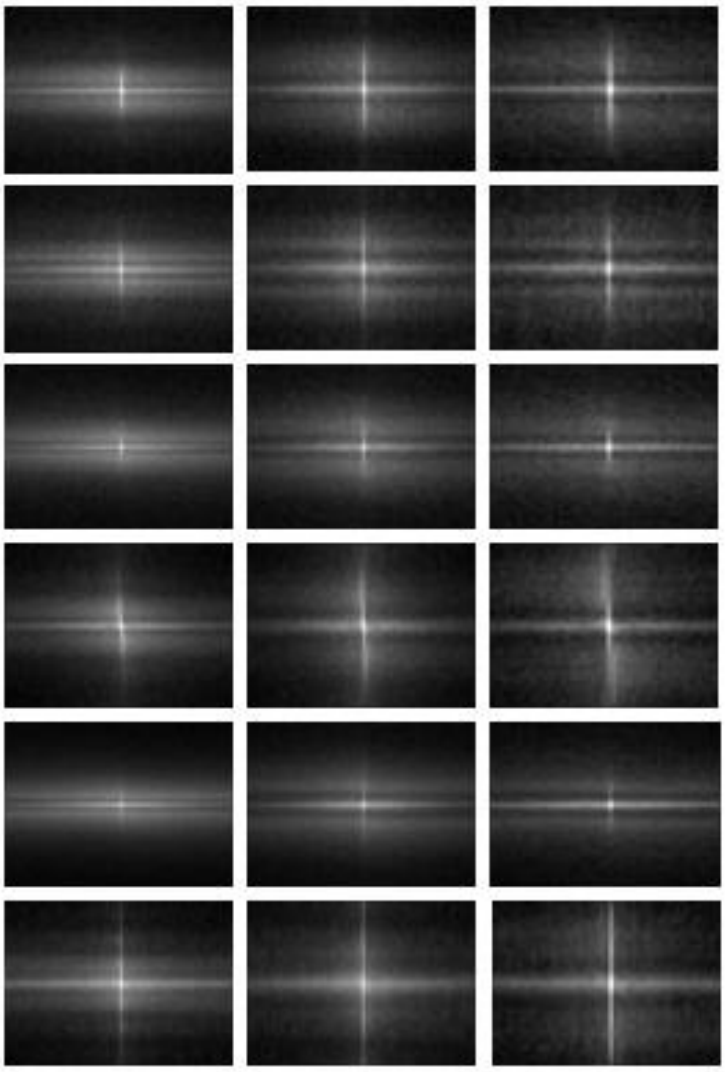
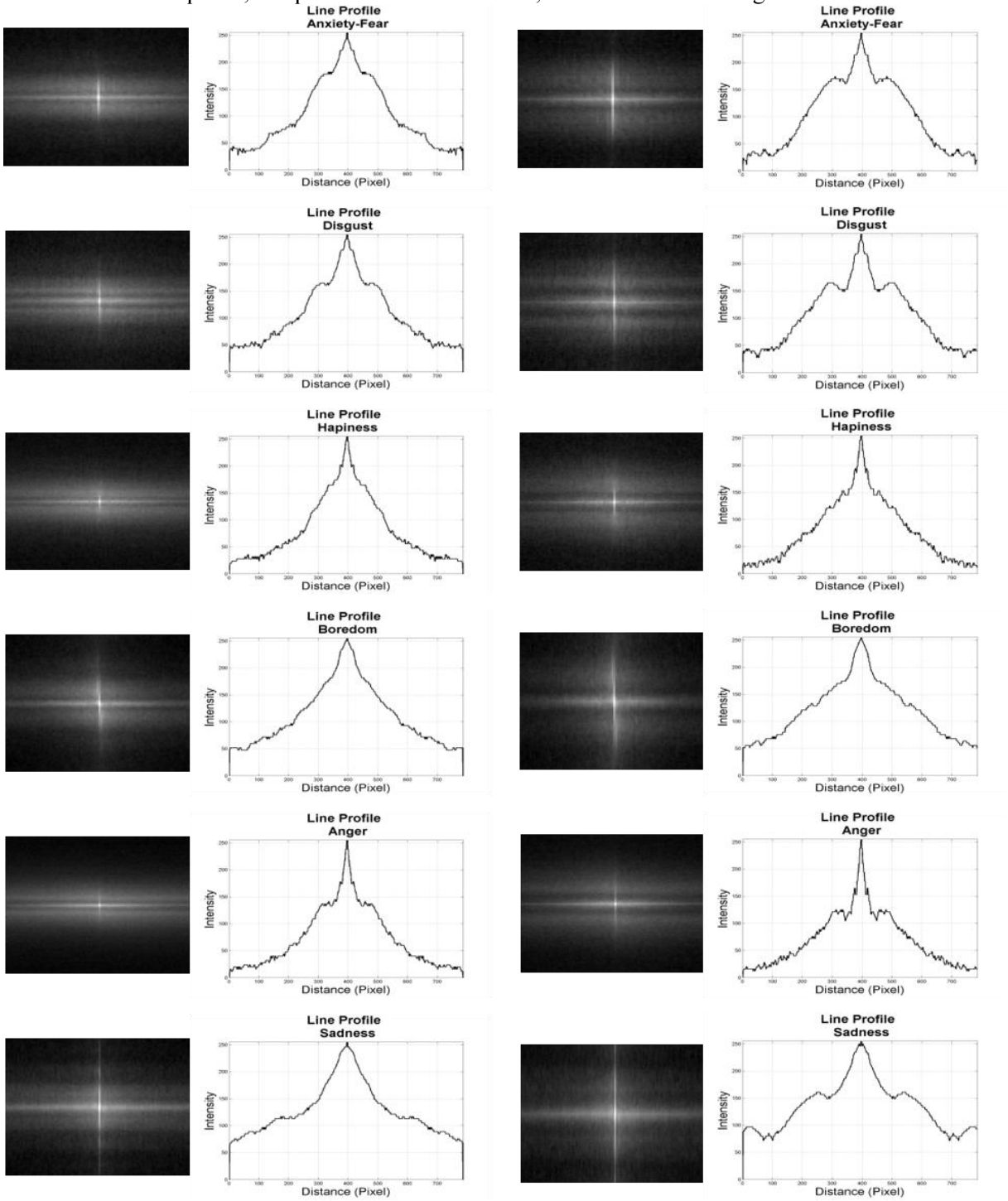


Figure 3. Spatial power spectral distributions and the corresponding vertical line profile for each emotional voiced speech, sampled at 8 KHz on the left, and 16 KHz on the right column



4. Conclusion

Evidences that emotions can be recognized when conducting a Double Fourier Analysis were presented. Spectrograms from voiced-speech recordings were obtained by using the short time Fourier transform, each of these time-frequency representations were considered as images, then from 2-

dimensional Fourier Transforms, obtained for each particular case, a spatial frequency distributions analysis were conducted. Particular structures related to emotion-induced features in voiced speech, become apparent in spatial power spectral density distributions, when synthesized by proposed method. Those structures are preserved even when signals are sampled at different frequencies, and different Gaussian windows are employed in the STFT to obtain spectrograms.
We are currently working on emotion classification when using Double Fourier Analysis, looking forward to extract and recognize emotional content in voiced speech, taking into advantage the particular features portrayed in spatial power spectra density functions.

Acknowledgement
Authors acknowledges Universidad EAFIT (Colombia) and Mathematical Modeling Research Group, where this research was performed.

References

[1] Borst, J.M. 1956. *Journal of the Audio Engineering Society* 14-23.
[2] Bolt, R.H., F.S. Cooper, David, E.E., Denes, P.B., Pickett, J.M., Stevens, K.N 1969 *Science*, **166** 338-342.
[3] Bolt, R. H., Cooper, F. S., David Jr, E. E., Denes, P. B., Pickett, J. M., & Stevens, K. N 1970 *The Journal of the Acoustical Society of America,* **47(2B)** 597-612.
[4] Mergu, R. R., & Dixit, S. K. 2011 *International Journal of Computer Applications*, **15(4)** 28-32.
[5] Oppenheim, A. V. 1970 *IEEE*, **7(8)** 57-62.
[6] Benesty, J., Sondhi, M. M., & Huang, Y. 2008 *Springer Science & Business Media*
[7] Pinkowski, B. "Principal component analysis of speech spectrogram images". Pattern recognition, 30(5), 1997, pp. 777-787.
[8] Welling, L., & Ney, H. 1998 *Speech and Audio Processing, IEEE Transactions on*, **6(1)** 36-48.
[9] Al-Darkazali, M., Young, R., Chatwin, C., & Birch, P. 2013 *In SPIE Defense, Security, and Sensing. International Society for Optics and Photonics*. 87480G-87480G
[10] Cadore, J., Gallardo-Antolín, A., & Peláez-Moreno, C. 2011 *In Advances in Nonlinear Speech Processing. Springer Berlin Heidelberg* 224-231
[11] Cadore, J., Valverde-Albacete, F. J., Gallardo-Antolín, A., & Peláez-Moreno, C. *Cognitive Computation* **5(4)** 426-44.
[12] Jin, X. C., & Wang, Z. F.2006 *18th International Conference on Pattern Recognition* **4** 278-281.
[13] Dennis, J., Tran, H. D., & Li, H 2011 *Signal Processing Letters, IEEE,* **18(2)** 130-133.
[14] Nilufar, S., Ray, N., Molla, M. K. I., & Hirose, K. 2012 *IEEE International Conference on Speech and Signal Processing (ICASSP),* 501-504
[15] Caelli, T., & Reye, D.1993 *Pattern recognition*, **26(4)** 461-470.
[16] Wang, K. C. The Feature Extraction Based on Texture Image Information for Emotion Sensing in Speech. Sensors, 14(9), 2014, pp. 16692-16714.
[17] Přibil, J., & Přibilová, A. 2010 *Measurement Science Review*, **10(3)** 72-77.
[18] Prandoni, P., & Vetterli, M. 2008. *Signal processing for communications. CRC Press.*
[19] Otsu, N. 1975 *Automatica* 23-27.
[20] Dennis, J., Tran, H. D., & Li, H. 2011 *Signal Processing Letters, IEEE*, **18(2)** 130-133.