TIME-FREQUENCY SIGNAL DECOMPOSITIONS FOR AUDIO AND SPEECH PROCESSING

Karthikeyan Umapathy[1] and Sridhar Krishnan[2]
Dept. of Elec. and Comp. Engg., [1]The University of Western Ontario, London, ON N6A 5B8 kumapath@uwo.ca
[2]Ryerson University, Toronto, ON M5B 2K3 krishnan@ee.ryerson.ca

1. INTRODUCTION

Efficient analysis and processing of audio signals would lead to a better utilization of computer vision and machine learning technologies in automating audio related applications. Audio and speech are highly non-stationary signals with a time-varying spectrum. It is difficult to analyze them using simple signal processing tools. Most of the existing techniques segment the audio signals and assume the signal to be quasi stationary within the short periods and apply stationary signal processing tools. However these approaches suffer from fixed time-frequency resolution and cannot accurately model the time varying characteristics of the audio signals. An adaptive joint time-frequency (TF) approach would be the best way to analyze audio signals.

The two well-known time-frequency approaches are based on 1. Signal decomposition, and 2. Bilinear TF distributions (also known as Cohen's class) [1]. In order to perform an objective analysis and to extract useful parametric information, the TF decomposition based approach would be ideal. Hence, the proposed methodology uses an adaptive TF transform (ATFT) based on the matching pursuit (MP) algorithm with Gaussian TF functions [2].

Majority of the audio and speech applications perform some combination of the following operations: (i) Compression and (ii) Feature extraction (for pattern recognition), and (iii) Denoising. The output specification for each of the above operations is grossly different. This paper is an attempt to present the proposed adaptive TF technique as a unified methodology (block diagram shown in Fig. 1) in addressing all the above operations on audio and speech signals. The paper is organized as follows: Section 2 covers the methodology comprising the subsections of ATFT, audio compression and audio & speech classification. Section 3 covers the time-width versus frequency band mappings. Discussion and Conclusions are given in Section 4.

2. METHODOLOGY

2.1 Adaptive time-frequency transformation

The core of the proposed methodology lies in the adaptive TF transformation based on the MP algorithm. MP, when used with a dictionary of TF functions yields an adaptive time-frequency transformation [2]. In MP any signal $x(t)$ is decomposed into a linear combination of K TF functions selected from a redundant dictionary of TF functions $g(t)$ as given by

$$x(t) = \sum_{n=0}^{K-1} \frac{a_n}{\sqrt{s_n}} g\left(\frac{t-p_n}{s_n}\right) \exp\left\{j(2\pi f_n t + \phi_n)\right\}$$ where a_n is

the expansion coefficient, the scale factor s_n also called as octave or time-width parameter is used to control the width of the window function, and the parameter p_n controls the temporal placement. The parameters f_n and ϕ_n are the frequency and phase of the exponential function respectively. The signal is projected over a redundant dictionary of TF functions with all possible combinations of scaling, translations and modulations. At each iteration, the best-correlated TF functions to the local signal structures are selected from the dictionary. The remaining signal called the residue is further decomposed in the same way subdividing them into TF functions.

2.2 Audio compression

In the audio compression application, we first modeled the audio signal (5s segments at 44.1k/s) with K number of TF functions that either captures 99.5 % of the signal energy or to a maximum of K=10,000. The TF decomposition parameters ($a_n, s_n, p_n, f_n,$ $\& \phi_n$) were analyzed and a novel TF psychoacoustics model was applied to discard the perceptually irrelevant TF functions. The perceptually filtered K^*TF functions were then quantized using 54 bits/ TF function. A curve fitting technique was used on the energy a_n parameter to significantly further reduce the total number of bits. An audio database containing 8 stereo signals of 20s long were used for testing. Compression ratios as high as 40 were achieved with an average SDG (subjective difference grade) of -1.1. The proposed technique performed exceedingly well for classical type of music compared to the existing techniques.

2.3 Audio and Speech Classification

A database of 170 audio signals containing 6 groups (rock, classical, country, folk, jazz and pop) of music

Thanks to NSERC and MICRONET for funding this research work.

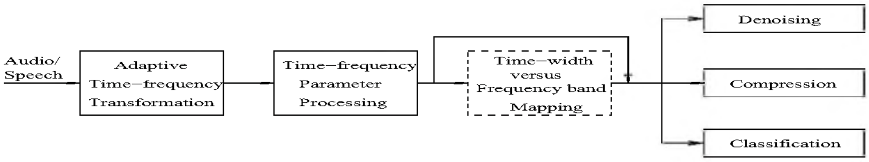


Fig. 1: Block diagram of the proposed methodology

signals were decomposed using the ATFT algorithm. The octave parameter s_n demonstrated high discrimination between the classes of signals. The octave distribution was computed over three frequency bands and used as features. A set of 42 features were extracted and used in classifying the six music groups. A classification accuracy of 97.6 % was achieved [3].

A database of 212 speech signals containing 51 normal and 161 pathological signals were used in the study. The distribution of octave parameter, the energy a_n capture rate and center frequency f_n of the TF functions demonstrated high discriminatory behavior between the normal and pathological signals. 3 features were derived and used for classifying the normal and pathological signals. A classification accuracy of 93.4% was achieved [4]. A sample energy a_n capture curve and octave s_n distribution are shown in Figs. 2(a) and 2(b). ATFT's inherent capability of denoising [2] helped in both the above discussed compression and classification applications to remove the insignificant signal components.

3. TIME-WIDTH VS FREQUENCY BAND MAPPING (TWFB)

Often when TF visualization of a signal is required, a TFD is constructed. Classical TFDs are non parametric and lacks the flexibility to relate visual patterns with a model or decomposition parameters. A more ideal choice of visualization would be that which preserves the parametric benefits of the decomposition. The idea is to generate a TF subspace mapping using the decomposition parameters that would serve as a (1) good parametric visualization tool, (2) an organized subspace mapping and (3) flexible TF subspace extractor that could address various denoising/ source separation problems. Of the five TF decomposition parameters discussed, (a_n, s_n & f_n) would be more appropriate to generate a subspace mapping. We form a 3D visualization by accumulating the energy for every combination (tile) of (s_n & f_n). The step size of (s_n & f_n) decides the resolution of the visualization. Figs. 2(c) to 2(f) show the spectrogram & TWFB mappings of a clean speech signal and a noisy speech signal (AWGN at 5dB). One can clearly see from Fig. 2(f), the tiles corresponding to the noise standing out separately. TWFB map is sensitive to signal structures; hence we could easily filter out or separate signal components that differ in structural content. This ability to segregate structural signal components can also be used to characterize different classes of (audio) signals.

4. DISCUSSION AND CONCLUSIONS

A unified adaptive TF decomposition based methodology for processing audio and speech signals was presented. The proposed non-stationary signal analysis tool performed well with diverse operations related to audio and speech applications. The proposed methodology is computationally expensive but considering the rapid hardware advancements this should not pose a problem in near future. A novel TWFB mapping was introduced which demonstrates high potential to form as a versatile parametric visualization/pattern recognition tool.

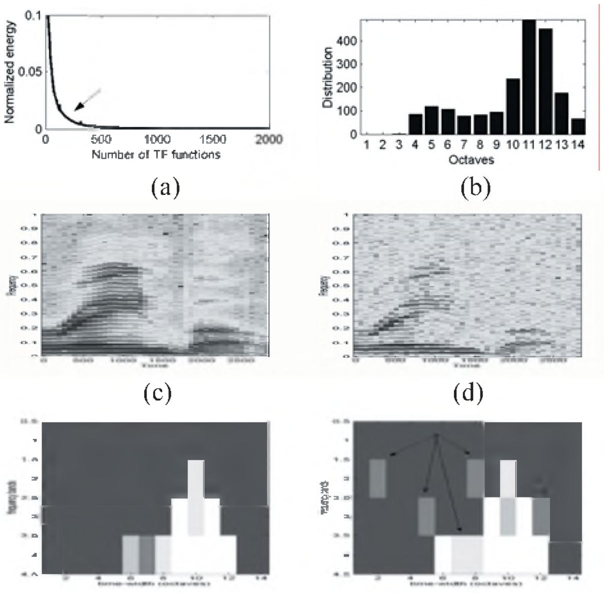
REFERENCES

[1] Cohen, L (1989). Time-frequency distributions – a review. Proceedings of the IEEE., vol 77, no 7, pp 941-981.

[2] Mallat, S . G. and Zhang. Z (1993). Matching pursuits with time-frequency dictionaries. IEEE Transactions on signal processing, vol 41, no 12, pp 3397-3415.

[3] Umapathy, K., Krishnan, S. and Jimaa, S. (2005). Multigroup classification of audio signals using time-frequency parameters. IEEE Transactions on Multimedia, vol 7, no 2, pp 308-315.

[4] Umapathy, K., Krishnan, S., Parsa, V. and Jamieson, D. G. (2005) Discrimination of pathological voices using a time-frequency approach. IEEE Transactions on Biomedical Engineering, vol 52, no 3, pp 421-430



(a) (b)

(c) (d)

(e) (f)

Fig. 2: (a) A sample energy capture curve, (b) A sample octave distribution, (c) Spectrogram of clean speech, (d) Spectrogram of a noisy speech (5dB), (e) TWFB map of a clean speech, (f) TWFB map of a noisy speech signal (5dB) showing distinctly the tiles corresponding to noise signal components.