AM-FM MODULATION FEATURES FOR MUSIC INSTRUMENT SIGNAL ANALYSIS AND RECOGNITION

Athanasia Zlatintsi and Petros Maragos

School of Electr. & Comp. Enginr., National Technical University of Athens, 15773 Athens, Greece

[nzlat,maragos]@cs.ntua.gr

ABSTRACT

In this paper, we explore a nonlinear AM-FM model to extract alternative features for music instrument recognition tasks. Amplitude and frequency micro-modulations are measured in musical signals and are employed to model the existing information. The features used are the multiband mean instantaneous amplitude (mean-IAM) and mean instantaneous frequency (mean-IFM) modulation. The instantaneous features are estimated using the multiband Gabor Energy Separation Algorithm (Gabor-ESA). An alternative method, the iterative-ESA is also explored; and initial experimentation shows that it could be used to estimate the harmonic content of a tone. The Gabor-ESA is evaluated against and in combination with Mel frequency cepstrum coefficients (MFCCs) using both static and dynamic classifiers. The method used in this paper has proven to be able to extract the fine-structured modulations of music signals; further, it has shown to be promising for recognition tasks accomplishing an error rate reduction up to 60% for the best recognition case combined with MFCCs.

Index Terms— AM-FM modulations, energy separation algorithm, music processing, timbre classification.

1. INTRODUCTION

Psychophysical research has shown that human hearing is largely based on amplitude and frequency modulations. The human auditory system through the transduction procedure, using the spectral shapes of auditory filters (FM to AM transduction), can perceive the frequency modulations [1, 2] of sounds. The musical signals' temporal microstructure consists of instantaneous amplitude and frequency modulations of their main resonances, which characterize the waveforms of those sounds. Modulations, such as the vibrato (FM) and the tremolo (AM), are easily understood, while smaller ones are not, nevertheless contributing to the creation of "natural" sounds [3], with particular importance in music composition [4]. Additionally, modulation analysis could be applied in the analysis of medium- and macro-structures for the description of different musical phenomena and the relations of their basic construction units.

Based on indications for the existence of nonlinear phenomena, i.e., modulations during speech production [5], such ideas have been used for speech analysis and especially for detection and recognition tasks. Maragos et. al [5] has proposed an AM-FM modulation model for speech and developed a nonlinear *Energy Separation Algorithm* (ESA) for demodulation of speech resonances in their amplitude and frequency components using bandpass filtering [6, 7, 8]. This kind of modeling has been used in applications of automatic speech recognition [9] and synthesis [10], while it has also been proved useful in speech recognition and detection in noisy conditions [9, 11].

Modulations have also been studied in [12] for the analysis and resynthesis of musical instrument sounds, in order to determine the synthesis parameters for an excitation/filter model. Similar ideas have been applied for recognition and specifically the distinction of speech and music in audio signals [13, 14, 15]. In [16], amplitude modulation features have been extracted as a set of features for instrument recognition so as to describe the tremolo measured in a frequency range between 4-8 Hz and the "roughness" of the played notes when the range is between 10-40 Hz. Similar ideas, based on a sinusoidal model [17], have been used for sound modeling [18] and source separation [19]. The main difference of the AM-FM model in comparison to sinusoidal model is that the latter does not have significant FM components apart from the frame to frame slow variation of the phase and the number of its components is almost one-order of magnitude larger than that of the modulation model which represents resonance components instead of harmonics.

In this paper, the analysis concerns isolated musical instrument tones, derived from the UIOWA database with instrument samples [20]. In Section 2, we motivate and explore the micro-modulations of musical signals, based on AM-FM modeling using the Gabor-ESA [9] for the demodulation. Additionally, we apply the iterative-ESA [7], for the estimation of the center frequencies f_c of the Gabor filterbank. In Sec. 3, we continue with recognition experiments, in order to exam-

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

ine the discriminability capabilities of the modulation features regarding instrument classification tasks, using both static and dynamic classifiers. We compare the descriptiveness of the extracted features against and in combination with a standard feature set of MFCCs and finally, we report on promising results regarding the AM-FM model used in this paper.

2. AMPLITUDE AND FREQUENCY MODULATION

Small fluctuations or micro-modulations in frequency occur naturally in both human voice and musical instruments. According to Bregman [21], such fluctuations are often very small, ranging from less than 1% for a clarinet tone to about 1% for a voice trying to hold a steady pitch, with larger excursions of as much as 20% for the vibrato of a singer. Bregman also states that even smaller amounts of frequency fluctuation could actually have important effects on the perceptual grouping of the existing component harmonics of a sound.

Herein, we assume that the musical signal can be represented as a combination of different "resonances", which approximately correspond to oscillation systems formed by the instruments' characteristics and the sound production procedure (e.g., instruments geometry, material, performance of a musical piece). Hence, certain frequencies are enhanced while others are reduced.

Inspired by similar ideas used for speech processing [5], we propose the modeling of each resonance component of music signals as an amplitude and frequency modulated sinusoid (AM-FM signal) while we model the whole music signal as a sum of such AM-FM components

$$S(t) = \sum_{i=1}^{K} \alpha_i(t) \cos\left(\phi_i(t)\right) \tag{1}$$

where α_i and ϕ_i are the instantaneous amplitude and phase signals of component i.

In each AM-FM signal, the instantaneous frequency models the time-varying frequency of the resonance, while the instantaneous amplitude follows the time-varying energy of the sound source producing the resonance. This model may estimate the average value of the frequency, the instantaneous amplitude of the resonance, and the instantaneous deviation of the frequency. The advantage of such an analysis is that AM-FM modulations are able to capture the fine structure and the rapid fluctuations of musical signals. Such modeling may be applied to smaller or larger analysis windows by exploring the modeling possibility of musical characteristics and theirs micro-, medium- and macro-structures.

2.1. Modulations Features

The AM-FM related features investigated in this paper are: the **mean Instantaneous Amplitude** (m-IAM) which is defined as the short-time mean of the instantaneous amplitude signal $|\alpha_i(t)|$ for each resonance component i, parameterizing the resonance amplitudes and capturing part of the nonlinear behavior of the signal, and the **mean Instantaneous Frequency** (m-IFM), which is a short-time weighted mean of the instantaneous frequency signal $f_i(t)$, which provides information about the signal's fine structure taking advantage of the excellent time resolution of the continuous-time ESA proposed by Maragos et al. [5].

The Energy Separation Algorithm (ESA), which makes use of the Teager Energy Operator [22] estimates the instantaneous amplitude and frequency signals given by

$$f(t) \approx \frac{1}{2\pi} \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \tag{2}$$

$$|\alpha(t)| \approx \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \tag{3}$$

where $\Psi[x] = \dot{x}^2 - x\ddot{x}$ and $\dot{x} = dx/dt$.

In this paper we use a regularized version of the ESA, called Gabor-ESA and proposed in [9], which is a combination of the continuous time ESA and Gabor filtering of the signal. Prior to the extraction of the features a Gabor filterbank consisting of twelve filters is applied to decompose the signal into bandpass components. The Gabor filters were chosen for their good joint time-frequency resolution [5]. In the frequency domain the filters were placed according to mel-scale with a bandwidth overlap of adjacent filters equal to 50%.

The Gabor-ESA gives smoother instantaneous estimates. In this case the operator Ψ and the bandpass filtering are combined as follows:

$$\begin{aligned} \Psi[x(t) * g(t)] = &\left[x(t) * \frac{dg(t)}{dt}\right]^2 \\ &- \left(x(t) * g(t)\right) \left[x(t) * \frac{d^2 g(t)}{dt^2}\right] \end{aligned} \tag{4}$$

where $x(t)$ is the input signal, and $g(t)$ is the Gabor impulse response.

2.2. Iterative-ESA for Estimating Filterbank Center Frequencies

In this section, we apply an alternative method for the estimation of the center frequencies f_c of the Gabor filterbank, the iterative-ESA [7]. This method implies the iterative application of ESA to the Gabor filtered signal and thus adjusting the center frequency of each filter after every iteration. The method is considered important since it reduces the importance of having good initial estimates of the center frequencies of the filterbank. For this analysis, we calculated the short-time instantaneous frequency of tones using 30 ms segments. Some of the tones used were A3 and A4 with fundamental frequency equal to 220 Hz and 440 Hz respectively

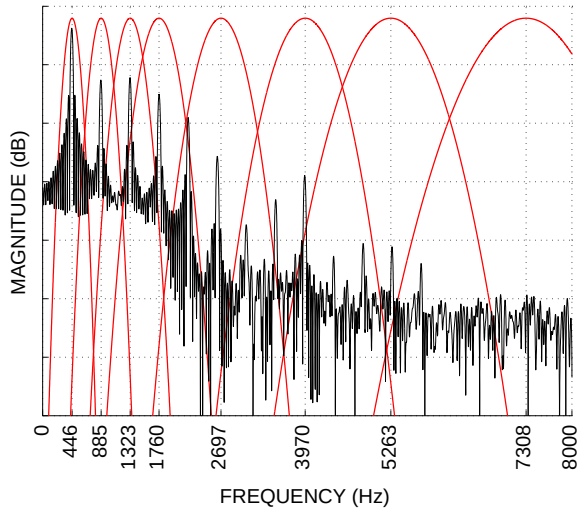

Fig. 1: Gabor filterbank with the estimated center frequencies f_c after the application of the iterative-ESA superimposed over Bb Clarinet spectrum for a 30 ms frame of the note A4, $Fs = 44.1$ Hz.

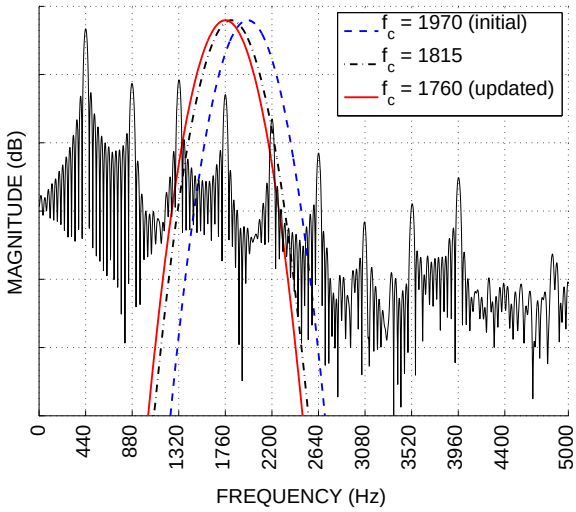

Fig. 2: Gabor filters superimposed over Bb Clarinet spectrum for a 30 ms frame of the note A4. The iterative-ESA for the fifth Gabor filter started at f_c = 1970 Hz and after two iterations converged to f_c = 1760 Hz which is $4f_0$ for the note A4, with a difference of 210 Hz.

from the instruments Bb Clarinet, Soprano Saxophone, Violin and Flute. We started the procedure using center frequencies dictated by the mel-scale, updating each one of them after every iteration of the ESA, while keeping the bandwidth fixed. The algorithm is assumed to have converged when the center frequency of each filter does not change by more than 1% or reached a certain number of iterations. Convergence was accomplished at average after four iterations for the low frequency filters, while we marked that more iterations were needed for high frequency filters.

The analysis showed that during this procedure the center frequencies tend to converge on frequencies which are close to integer multiples of the fundamental frequency of the analysis tone, i.e., the harmonics. Figure 1 shows the Gabor filterbank with the updated estimates of the center frequencies f_c superimposed over the spectrum of a 30 ms analysis frame of the note A4 (f_0: 440 Hz) of Bb Clarinet. The frequencies shown on x-axis are the estimated frequencies for the filter numbers two through nine and the signal is shown up to 8 kHz. As seen, these frequencies are actually close estimates to f_0, $2f_0$, $3f_0$, $4f_0$, $6f_0$, $9f_0$, $12f_0$, and $17f_0$. In Fig. 2, the procedure of convergence can be seen for the fifth Gabor filter, superimposed over the spectrum of a 30 ms segment for the note A4 from Bb Clarinet. The initial frequency was equal to 1970 Hz while after two iterations it converged to f_c = 1760 Hz which is actually the fourth harmonic ($4f_0$) of the note A4. Similar results were gained from the analysis of the other instruments too. Another important observation was that some of the Gabor filters favored to converge at the same center frequency. This is something that remains to be explored to find out whether it is due to the initially chosen center frequencies or to the signal's properties at these frequencies where there are no accentuated harmonics. However, we assume that our findings are significant and require further exploration since they gave us strong evidence that such a method could produce better estimates of $|\alpha(t)|$ and $f(t)$ while it shows a certain ability to estimate the harmonic content of the tone, despite the fact that there is no prior knowledge of the examined tone.

3. RECOGNITION EXPERIMENTS

In this section, we investigate the recognition properties of the proposed features. Two sets of experiments were carried out: (1) 1331 notes were used from seven different instruments, which are Double Bass, Bassoon, Cello, Bb Clarinet, Flute, Horn and Tuba; (2) five more instruments (738 notes) were added, and they are Alto Saxophone, Bass Trombone, Tenor Trombone, Bb Trumpet and Oboe, thus a total of 12 instruments were used to evaluate the features. For both sets of experiments, same parameters were used. The collection consists of the instruments' full range for the dynamic range piano to forte and the signals are sampled at 44.1 kHz. The different cases of feature sets were evaluated using static (GMMs) and dynamic classifiers (HMMs) in order to model the temporal characteristics of the signals too. The experimentation consisted of diverse combinations of N [3-9] states and M [1-3] mixtures. For the implementation of the Markov models the HTK [23] HMM-recognition system was used, by means of EM estimation using the Viterbi algorithm, adopting a left-right topology for the modeling. The results obtained are after five-fold cross validation with randomly se-

Feature Sets	
Proposed Features	
1	AMFM (12 m-IAM + 12 m-IFM)
2	$AMFM_{\Delta}$ (12 m-IAM+12 m-IFM (+ their 12 Δ + 12 $\Delta\Delta$))
3	$AMFM_{50}$ (50 AMFM features after PCA)
4	$AMFM_{39}$ (39 AMFM features after PCA)
5	$MFCC_{\Delta}$ (13 MFCC + 13 Δ + 13 $\Delta\Delta$)
Multi-Stream Cases	
1	$AMFM_{\Delta}$ + $MFCC_{\Delta}$
2	$AMFM_{39}$ + $MFCC_{\Delta}$
3	$AMFM_{50}$ + $MFCC_{\Delta}$

Table 1*:* List of feature sets used in recognition experiments.

lected training set, using 70% of the available tones. The ability of the examined features was further compared to a standard feature set of 13 MFCCs (with their first and second temporal derivatives), which are chosen both for their good performance and the acceptance they have gained for instrument recognition tasks. The analysis of the MFCCs was performed in 30 ms windowed frames with a 15 ms overlap, and with 24 triangular bandpass filters. For the combined feature sets, a multi-stream configuration was adopted where each subset of features was trained in a different stream and then fused employing different stream weights for experimentation purposes.

In our experiments, we evaluate the performance of fixed sets of features, which are listed in Table 1. The mean-IAM and mean-IFM features are estimated in 30 ms frames with a 15 ms overlap. For the demodulation, twelve Gabor filters were used since it was empirically found to be a good choice, after extensive experimentation. The first and second temporal derivatives of the features were extracted resulting in a 72 $AMFM_{\Delta}$ feature vector. The dimensionality of the $AMFM_{\Delta}$ feature space, consisting of the mean-IAM and mean-IFM, was reduced using PCA, in order to decorrelate the data and obtain the optimal number of features that accounts for the maximal variance. Several different combinations of the number of PCA components were examined in order to investigate how the discriminability results varied and thus were enhanced. The study showed that the mean-IFM features were better decorrelated thus more were needed in order to obtain the maximum discriminability among the examined instruments. The cases presented next, accomplished the higher error reduction compared to $MFCC_{\Delta}$ and to the full set of $AMFM_{\Delta}$ features. In the first case, the reduced feature space of total 50 PCA components consists of 18 mean-IAM components (6 m-IAM, 6 m-IAM_{Δ}, 6 m-$IAM_{\Delta\Delta}$) and 32 mean-IFM (12 m-IFM, 10 m-IFM_{Δ}, 10 m-$IFM_{\Delta\Delta}$). Since our intentions were to acquire as small as possible feature space or at least comparable in number with the 39 $MFCC_{\Delta}$, we reduced the principal components to 39 using 12 mean-IAM components (4 m-IAM, 4 m-IAM_{Δ}, 4 m-$IAM_{\Delta\Delta}$) and 27 mean-IFM (12 m-IFM, 8 m-IFM_{Δ}, 7 m-$IFM_{\Delta\Delta}$).

Accuracy Results for 7 Instruments				
Feature Set	Weights	GMM	HMM	
Proposed Features				
	MFCC-AMFM	$M=3$	$N=3$ $M=3$	$N=5$ $M=3$
AMFM	-	88.74	94.90	95.00
$AMFM_{\Delta}$	-	89.14	94.60	96.72
$AMFM_{50}$	-	**95.30**	96.31	96.06
$AMFM_{39}$	-	94.29	**96.41**	**96.77**
$MFCC_{\Delta}$	-	86.06	94.65	96.16
Multi-Stream Cases				
$MFCC_{\Delta}$ - $AMFM_{\Delta}$	1.00 - 0.50	**91.57**	**96.67**	97.57
	0.50 - 1.00	90.50	95.66	97.17
	0.50 - 0.50	90.91	96.61	**97.93**
	1.00 - 0.10	90.10	96.52	96.97
$MFCC_{\Delta}$ - $AMFM_{50}$	1.00 - 0.50	95.81	**98.33**	**98.64**
	0.50 - 1.00	**96.26**	97.78	97.67
	0.50 - 0.50	**96.26**	97.88	97.83
	1.00 - 0.10	89.6	96.46	97.73
$MFCC_{\Delta}$ - $AMFM_{39}$	1.00 - 0.50	95.46	**98.48**	**98.68**
	0.50 - 1.00	**96.31**	97.82	98.13
	0.50 - 0.50	96.16	98.18	98.18
	1.00 - 0.10	90.71	96.61	97.37
Accuracy Results for 12 Instruments				
Proposed Features				
$AMFM_{50}$	-	**85.46**	91.74	93.72
$AMFM_{39}$	-	82.38	**92.68**	**93.72**
$MFCC_{\Delta}$	-	79.09	88.23	90.60
Multi-Stream Cases				
$MFCC_{\Delta}$ - $AMFM_{50}$	1.00 - 0.50	88.55	94.37	95.32
	0.50 - 1.00	87.19	**94.73**	94.96
	0.50 - 0.50	**88.03**	94.50	**95.55**
	1.00 - 0.10	86.18	92.33	94.02
$MFCC_{\Delta}$ - $AMFM_{39}$	1.00 - 0.50	**87.64**	**95.19**	**95.89**
	0.50 - 1.00	85.33	94.67	95.45
	0.50 - 0.50	86.70	94.93	95.67
	1.00 - 0.10	85.73	92.55	93.33

Table 2*:* Recognition accuracy results for 7 and 12 instruments, where N denotes the number of states and M the number of mixtures. For feature set specific information, see Table 1.

3.1. Results

The obtained accuracy scores of the classification results for the different cases of feature sets were promising and proved out to yield better recognition than the $MFCC_{\Delta}$ for most cases (even those not presented here). The most representative for both sets of experiments are reported in Table 2.

We notice that $AMFM_{\Delta}$ showed higher discriminability than the $MFCC_{\Delta}$ with an error reduction of 15% for N = 5, M = 3. The best case of $AMFM_{39}$ yields and error reduction up to 60% (15%), 33% (38%) and 16% (33%) for the GMMs and the HMMs when N = 3, 5 and M = 3, for 7 and 12 instruments (the error reduction for 12 instruments can be seen in brackets). We herein assume that the AMFM features are favorable and they accomplish correct recognition among the analyzed instruments.

The combination of the proposed features ($AMFM_{39}$) with the $MFCC_{\Delta}$ is acquiring even higher error rate reduc-

tion, which is ca. 60% and 56% in comparison to the MFCC_{Δ} for 7 and 12 instruments respectively. The scores regarding the stream weights that were used for the experimentation are comparable, with slightly better being the case were the weights are set equal to $s_1 = 1.00$ for the MFCC_{Δ} and $s_2 = 0.50$ for the various AMFM feature sets. However, for the case where the MFCC_{Δ} stream weight is equal to $s_1 = 1.00$ and for the AMFM features $s_2 = 0.10$, we mark that the obtained accuracy is much lower, which strengthens the fact that the AMFM features contribute remarkably in the recognition task. Furthermore, we notice that HMMs receive greater results, since they imply the temporal information of the tones too, although the error reduction for the proposed features compared to MFCCs is higher for the classification cases using GMMs.

4. CONCLUSIONS

In this paper, we presented a nonlinear AM-FM model for the demodulation of musical signals to instantaneous amplitude and frequency modulation signals, motivated by similar successful ideas applied to speech recognition and speech/music discrimination tasks. One of our long term goals in this area is to gain insight about the instruments' properties. In this paper we examined the discriminability capabilities of the modulation features regarding instrument classification tasks. Based on the the evaluation scores from two sets of experiments, strong indications have arisen that modulation features can capture important aspects of music sounds and discriminate among different instruments.

On that account, in our ongoing research, we are applying the method to a full set of instruments to validate the results while increasing the difficulty of the recognition task by inserting more instruments of the same family. We would also like to improve our preliminary work on iterative-ESA and examine whether it could endorse our first observations and integrate it in the analysis of the micro-structure of the signals. Moreover, we plan to perform a more careful and complete analysis of the AM-FM model regarding the medium- and macro-structures of musical signals.

5. REFERENCES

[1] T. F. Quatieri, T. E. Hanna, and G. C. O'Leary, "AM-FM Separation using auditory-motivated filters," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 5, pp. 465–480, Sep. 1997.

[2] W. Torres and T. Quatieri, "Estimation of modulation based on FM-to-AM transduction: Two-sinusoid case," *IEEE Trans. Signal Processing*, vol. 47, no. 11, pp. 3084–3097, 1999.

[3] P. M. Warren, *Auditory Perception*, Cambridge University Press, 3rd edition, 2008.

[4] D. E. Hall, *Musical Acoustics*, Brooks/Cole, 3rd edition, 2002.

[5] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[6] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3245–3265, 1993.

[7] H. M. Hanson, P. Maragos, and A. Potamianos, "A system for finding speech formants and modulations via Energy Separation," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 436–443, July 1994.

[8] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Amer.*, vol. 9, no. 6, pp. 196–200, Jun. 1996.

[9] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, Sep. 2005.

[10] A. Potamianos and P. Maragos, "Speech processing applications using an AM-FM Modulation model," *Speech Communication*, vol. 28, pp. 195–209, 1999.

[11] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2024–2038, 2006.

[12] R.B. Sussman and M. Kahrs, "Analysis and resynthesis of musical instrument sounds using Energy Separation," in *Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP-96)*, 1996.

[13] S. C. Sekhar and T.V. Sreenivas, "Novel approach to AM-FM decomposition with applications to speech and music analysis," in *Int'l Conf. Acoustics, Speech, and Signal Processing*, 2004, vol. 2, pp. 753–756.

[14] S. K. Kopparapu, M. A Pandharipande, and G. Sita, "Music and vocal separation using multiband modulation based features," in *IEEE Symposium on Industrual Electronics and Applications*, 2010, pp. 733–737.

[15] O. M. Mubarak, E. Ambikairajah, J. Epps, and T. S. Gunawan, "Modulation features for speech and music classification," in *Int'l Conf. Communication systems (ICCS-2006)*, Oct. 2006, pp. 1–5.

[16] S. Essid, G. Richard, and B. David, "Musical instrument recogntion by pairwise classification strategies," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1401–1412, 2006.

[17] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[18] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, Pope S. Picialli A. Roads, C. and G. (Eds.) De Poli, Eds. Swets & Zeitlinger, 1997.

[19] J.J. Burred and T. Sikora, "Monaural source separation from musical mixtures based on time-frequency timbre models," in *Proc. Int'l. Conf. on Music Information Retrieval (ISMIR-07)*, 2007.

[20] University of Iowa Musical Instrument Sample Database, ," [ONLINE], Available: http://theremin.music.uiowa.edu/.

[21] A. S. Bregman, *Auditory Scene analysis, The perceptual organization of sound*, MIT Press: Cambridge, MA, 1990.

[22] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*, W.J. Hardcastle and A. Marchal, Eds., vol. 15. NATO Advanced Study Institute, Series D, Boston, MA: Kluwer, Jult 1989.

[23] S.Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book. Revised for HTK Version 3.2*, Cambridge Research Lab, Dec 2002.