ANALYSIS AND AUTOMATIC RECOGNITION OF HUMAN BEATBOX SOUNDS: A COMPARATIVE STUDY

Benjamin Picart[1], *Sandrine Brognaux*[1,2], *Stéphane Dupont*[1]

[1]TCTS Lab, Faculté Polytechnique (FPMs), University of Mons (UMons), Belgium
[2]Cental - ICTEAM, Université catholique de Louvain, Belgium

benjamin.picart@umons.ac.be, sandrine.brognaux@uclouvain.be, stephane.dupont@umons.ac.be

ABSTRACT

"Human BeatBox" (HBB) is a newly expanding contemporary singing style where the vocalist imitates drum beats percussive sounds as well as pitched musical instrument sounds. Drum sounds typically use a notation based on plosives and fricatives, and instrument sounds cover vocalisations that go beyond spoken language vowels. HBB hence constitutes an interesting use case for expanding techniques initially developed for speech processing, with the goal of automatically annotating performances as well as developing new sound effects dedicated to HBB performers. In this paper, we investigate three complementary aspects of HBB analysis: pitch tracking, onset detection, and automatic recognition of sounds and instruments. As a first step, a new high-quality HBB audio database has been recorded, carefully segmented and annotated manually to obtain a ground truth reference. Various pitch tracking and onset detection methods are then compared and assessed against this reference. Finally, Hidden Markov Models are evaluated, together with an exploration of their parameters space, for the automatic recognition of different types of sounds. This study exhibits very encouraging experimental results.

Index Terms— Human beatbox, pitch tracking, onset detection, Hidden Markov Model, automatic speech recognition

1. INTRODUCTION

Speech and voice have been an intensive research topic in the past, leading to various technologies for synthesizing speech, and recognizing words or speakers. Singing voice has been studied too, but to a much lesser extent. Other forms of vocalization are an almost untouched area. These include some traditional and popular vocal expression forms such as Corsican, Sardinian, Byzantine, or contemporary singing expressions such as Human BeatBox (HBB). These are currently studied in the i-Treasures project [1, 2]. In this study, we focus on HBB.

HBB is an interesting and challenging case study related to spoken language technologies. Indeed, notations proposed for drum sound imitations often rely on the international phonetic alphabet, and performers produce sounds that are close to stop consonants, although they rely on a larger set of variants. Besides, performers make use of both ingressive and egressive sounds, which is rather unusual for standard spoken language technologies.

In the past, various studies focused on HBB, e.g. on the description of acoustic properties of some sounds used in HBB compared to speech sounds [3], based on author's observations; the investigation of HBB as a query mechanism for music information retrieval [4]; and the automatic classification of HBB sounds amongst kick/hihat/snare categories [5, 6]. More recently, the repertoire of a human beatboxer was analyzed by real-time magnetic resonance imaging [7], where the articulatory phonetics involved in HBB performance were formally described. The vocal tract behaviour in HBB was analyzed through fiberscopic imaging [8], to understand how they manage instrumental, rhythmic and vocal sounds at the same time. However, to the best of our knowledge, there are no reported comparative evaluations of HBB pitch tracking, onset detection, and automatic recognition, to the level proposed in this paper. Also, studies of automatic recognition of HBB pitched instrument imitations are very scarce. These motivate the work reported here.

The paper is structured as follows. Section 2 presents the recording protocol and the content of our new high-quality HBB audio database. This database is carefully segmented and annotated manually to obtain a ground truth reference, against which various pitch tracking and onset detection methods are evaluated in Section 3. In Section 4, we investigate the use of Hidden Markov Models (HMMs) for the automatic recognition of different sound types present in our HBB database. Finally, Section 5 concludes the paper.

2. CREATION OF A DATABASE WITH VARIOUS HBB STYLES

For the purpose of this research work, a new beatbox database was recorded. It consists of sounds produced by two male beatboxers (Davox and Matthieu), spread over two non-consecutive sessions (session A: Davox and Matthieu; session B: Davox only). Each session contains 4 sets: individual drum sounds, instruments, rhythms and freestyle. The beatboxers were placed, one at a time, inside a soundproof room, equipped with a computer for recording, and with a Rode Podcaster microphone at a distance of approximately 20 cm from the performer's mouth. The audio was captured at a sampling rate of 48 kHz, and ElectroGlottoGraphy (EGG) signals were recorded for each beatboxer during session A. In total, we were able to collect more than 9000 musical events. In this study, we will use 1835 drum sounds and 1579 musical instrument notes imitation.

The first set of the database, i.e. individual drum sounds, can be divided into five main classes: cymbal, hihat, kick, rimshot and snare. For each of these, beatboxers have their own variants, imitating drum sound timbers typical to various music styles. For instance, Davox pronounced several time the 17 HBB drum sounds described in [7], while Matthieu proposed his own repertoire (the repetition amount is detailed in Table 1). Each onset of a drum sound event is manually annotated and labeled according to the beatboxer's repertoire. In this study, we did not try to recognize those variants, but only the five broad categories (see Section 4).

For the second set, they were asked to imitate the sound of various instruments, with custom rhythms and melodies for each (the

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

Table 1: *Musical classification and acoustic characteristics (and repetition number #) of drum sounds and HBB instruments.*

	Musical classification (#)	Acoustic characteristics (#)	
		Davox	Matthieu
Drums (#)	Hi-hat (436)	h (45), kss (32), t (18), th (32), tss (47)	t (71), th (47), ts (44), frr (44), sip (56)
	Cymbal (147)	kshh (32), tsh (29)	fsh (31), psh (29), soh (26)
	Kick (433)	bf (39), bi (34), bu (39)	pf (65), po (73), tu (65), voc (69), vocnas (49)
	Rimshot (525)	k (51), kh (39), khh (55), suckin (35)	k (59), k_ing (89), k_ingvoc (51), kh (46), ko (45), ks (55)
	Snare (294)	clap (17), pf (48), ksh (26)	ich (56), if (42), pf (59), plf (46)
Instruments (#)	Elec. guit. egr. (119)	no_effect (111), inhalation (5), sil (3)	*No performance*
	Elec. guit. ing. (510)	voiced (153), unvoiced (39), inhalation (30), sil (4)	voiced (154), intermod_dist (90), inhalation (12), sil (28)
	Guit. bass (148)	no_effect (102), inhalation (11), sil (0)	no_effect (19), inhalation (16)
	Saxo. (251)	no_effect (122), pre_breath (101), inhalation (20), sil (8)	*No performance*
	Trumpet (210)	no_effect (141), vibrato (15), inhalation (3), sil (4)	no_effect (39), tremolo (3), inhalation (1), sil (4)
	Trump. cork. (137)	no_effect (67), vibrato (9), inhalation (3), sil (6)	no_effect (42), inhalation (2), sil (8)
	Trump. trill. (136)	no_effect (64), tremolo (28), vibrato (3), inhalation_tss (3), inhalation (8), sil (30)	*No performance*
	Didgeridoo (49)	*No performance*	didgeridoo (39), inhalation (4), sil (6)
	Harmonica (19)	*No performance*	harmonica (9), vibrato (6), inhalation (4)

collected amount of notes is detailed in Table 1). Some performances actually contain different instrument timbres, or specific sounds produced to increase the performance realism (e.g. unvoiced sounds to simulate more muted guitar notes, exhalation sounds preceding saxophone notes, as well as vibrato and tremolo effects in trumpet imitations). This set has been manually segmented and labeled, according to its own characteristics. We also recorded some special instruments, which will not be used in this study: voice scratch (or the imitation of disc jokey turntable scratch) with Davox, electric guitar ingressive and guitar bass superimposed with beats with Matthieu. Here also, our experiments targeted the recognition of instrument categories rather than subtle variants proposed by the beatboxers.

3. HBB DATA ANALYSIS AND FEATURE EXTRACTION

This section aims at developing techniques enabling the extraction of relevant features characteristic of vocal performances: pitch of the notes and musical event onsets.

3.1. Pitch analysis

Extracting pitch makes sense for instrument imitations, as drum sounds do not contain any pitch. Similarly to [9], we compare the performance of 4 of the most representative state-of-the-art pitch extraction techniques: RAPT [10], SRH & SSH [11], YIN [12].

Usually, ground truth reference pitch for voice is obtained through EGG recordings. These signals were captured only during session A of our database. For both sessions A and B, the reference pitch was obtained by applying an automated pitch tracking algorithm (Praat [13]), followed by manual and thorough checks and corrections. Contrarily to standard neutral speech, we observed that EGG is not sufficient on its own to get accurate pitch ground truth.

It should be noted that all applied algorithms allow, in addition to pitch values extraction, Voiced/Unvoiced (VUV) decisions computation as a by-product. These two aspects of pitch extraction should be separately evaluated, so as to find the most appropriate method for each of them. As performance measures [14], we used Voicing Decision Error (VDE - proportion of frames with a voicing decision error), Gross Pitch Error (GPE - proportion of frames with relative F0 error higher than a 20% threshold) and F0 Frame Error (FFE - proportion of frames with either GPE or VDE error).

Table 2 summarizes the average pitch tracking results. Among the four algorithms mentioned above, the best scores are achieved by SRH for GPE, VDE and FFE. Major degradations contributing to increase these scores may come from electric guitar egressive from Davox, and from electric guitar ingressive from Matthieu. Indeed, we noted the presence of strong sub-harmonics (imitating intermodulation distortion, which is common in heavily distorted guitar sounds) in the corresponding audio signals, leading pitch tracking algorithms to fail selecting the correct fundamental.

Table 2: *Pitch tracking results, computed on around 72000 frames (around 11 minutes of audio), for the instrument imitations.*

	GPE [%]	VDE [%]	FFE [%]
RAPT	29.3	13.8	19.6
SRH	12.7	11.7	15.5
SSH	14.8	14.0	17.6
YIN	14.2	14.3	18.2
MAJORITY	**10.3**	**9.5**	**13.3**

Finally, we proposed to use majority voting approach to further improve pitch estimation accuracy. In case of a tie, pitch values and VUV decisions are set from SRH, as this algorithm led to the best scores compared to the others. This approach yields in significant performance improvement (as clearly observed in Table 2).

3.2. Onset detection

Onset detection is a specific problem within the area of audio analysis and recognition, and can be the first step in a system designed to interpret an audio stream in terms of its relevant audio/musical events. It is also particularly relevant to HBB due to its mostly percussive basis. It usually consists in three steps [15, 16]. First, the audio signal is pre-processed in order to accentuate certain important aspects, using techniques including filtering, separation into frequency bands, or the separation of transient and steady-state portions of the signal. Then, the amount of data from the processed signal is

Table 3: *Onset detection results for the drum sounds (1835 onsets, i.e. around 10 minutes and 30 seconds of audio) and instrument imitations (1579 onsets, i.e. around 11 minutes of audio).*

	Method	F-measure	Precision	Recall
Drums	S. Flux on Log Mag.	92.2%	92.4%	91.9%
	Weight. Phase Div.	92.1%	92.2%	91.9%
	Log Magnitude	91.0%	91.9%	91.9%
Instruments	Weight. S. Flux on Log Mag.	82.4%	81.6%	83.2%
	S. Flux on Log Mag	79.3%	79.3%	79.3%
	Kullback-Leibler div.	79.0%	78.3%	79.8%

reduced so as to obtain a lower sample rate onset detection function, where the onsets manifest themselves as peaks. Finally, thresholding and/or peak-picking can be applied to isolate potential onsets.

A well known approach consists in considering the signal high-frequency content by linearly weighting each bin in proportion to its frequency, hence emphasizing high-frequencies typical to discontinuities, and especially onsets [17]. A more general approach consists in considering changes in the spectrum and formulate detection functions as distances between consecutive short-term Fourier transforms (STFTs), such as "spectral flux" approach when distance is computed between successive power spectra. Discontinuities in the evolution of phase spectra, of complex spectra [18, 19] have also been shown beneficial.

3.2.1. Method

Comparative evaluations on instrument onset detection are available in the literature but not on HBB data. Here, evaluation is performed using an approach similar to [19], where a detected onset is considered as correct if it lies within a tolerance margin of 50 ms before and after the ground truth (hand annotated) onset.

The compared methods have been reimplemented by the authors, and cover a range of 18 variants of magnitude-based (including energy, log-energy domains, and their time derivatives) and STFT-based (including spectral flux in different domains, phase-based methods and their variants using magnitude weighting, and complex-based methods) approaches. We optimized the detection threshold for peak F-measure for each method. The number of sound event being limited, a k-fold method was applied (k = 3).

3.2.2. Results

Table 3 presents the results for the three best performing methods on drum sounds and instrument imitations.

On drums data, many approaches work very well. There is a significant error rate due to the miss of some onsets though, such as those of hi-hats *th* and *t*, which can be up to 24 dB weaker than kick or snare drums. Also, longer sounds such as cymbals can sometimes trigger two onsets, the second one due to modulations applied by the performer. The three best methods were: 1) a spectral flux approach [20] operating on the log-magnitude bins of the STFTs; 2) a STFTs phase divergence approach [19] where bin importance is weighted according to their magnitudes; 3) a method making use of the log-magnitude of signal frames along time (with no spectral flux computation). This represent a simple method that has, to our knowledge, not been proposed before.

On pitched instruments data, the performance level is lower, essentially due to modulations, false attacks, and onsets detected at the end of sounds. The three best methods were: 1) a spectral flux approach operating on the log-magnitude bins of the STFTs where bin importance is weighted according to equal loudness contours; 2) a spectral flux approach operating on the log-magnitude bins of the STFTs; 3) an approach based on the modified Kullback-Leibler distance on the log-magnitude bins of the STFTs [16].

Spectral flux computed on log-magnitude spectra appears as a good compromise, with F-measures reaching 92% and 79% on drums and pitched instruments respectively. These results are hence in line with those reported on real instrument sounds (see [19]).

4. AUTOMATIC HBB SOUNDS AND INSTRUMENTS RECOGNITION

Automatic recognition systems were initially developed for speech (ASR, e.g. [21, 22, 23]). In this section, we assess the performance of such systems using our HBB audio data for automatically annotating any new HBB performances. As already mentioned in Section 2, we chose to recognize the musical classification classes, rather than subtle timber variants (detailed in Table 1).

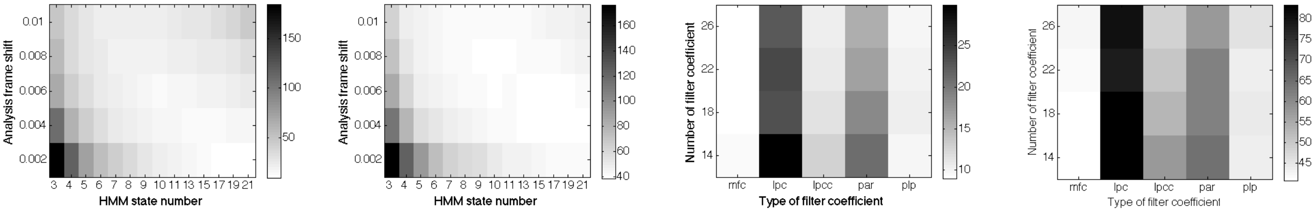
4.1. Method

HMM-based HBB recognizers were built relying on the implementation of the publicly available HTK toolkit [24] (version 3.4.1). Given the nature of the audio material (similar to spoken language, but with different dynamics), we founded necessary to re-explore the common assumptions made when HMMs are used in ASR. Among those, the audio analysis frame shift, the feature extraction approach and the HMM state number for each sound. Given the large parameters space, we proceeded in two steps: (i) joint exploration and optimization of the parameters related to temporal properties (analysis frame shift and HMM state number), (ii) selecting the best configuration from the previous step, joint exploration and optimization of feature extraction approach and number of feature coefficients. The HBB audio data being limited, the same k-fold method (k = 3) as in Section 3.2.1 was used on the first and second sets of the database (i.e. 1835 drum sounds and 1579 instrument imitations), for training and testing the HMM-based HBB recognizers.

For step (i), we extracted as filter parameterization the 54 Mel-Frequency Cepstral Coefficients (MFCCs) traditionally used in ASR[1]: 18 MFCCs and their first- and second-order derivatives. The length of the analysis window was equal to 10 ms, as individual drum sounds as short as 10 ms were observed. We varied the frame shift from 2 ms to 10 ms by 2 ms increments. Moreover, we also varied the number of states composing the HMMs from 3 to 11 by 1-state increments, and from 13 to 21 by 2-state increments. For each drums and pitched instruments type, a no-skip left-to-right single-Gaussian monophone HMM was built. Silence and short-pause models were also implemented, as proposed in [24]. Models initialization and training is based on our manually-aligned corpus.

For step (ii), feature extraction approach was analyzed. HMMs were trained using from 14 to 26 feature coefficients incremented by 4, with MFCCs, Linear Predictive Coding (LPC) coefficients, Linear Prediction Cepstral Coefficients (LPCC), PARtial CORrelation (PARCOR) and Perceptual Linear Prediction (PLP) coefficients [21].

[1] The energy coefficient was left aside and the audio recordings were not normalized. Indeed, we are targeting a real-time application and evaluation with actual end-users.



(a) Drums (b) Instruments (c) Drums (d) Instruments

Fig. 1: Automatic recognition of HBB drum sounds and instrument imitations - Evolution of the error rate: (a & b) as a function of frame shift and state number, (c & d) as a function of feature extraction approach and number of coefficients.

The performance scoring was obtained with the sclite tool, which is part of NIST Scoring Toolkit (SCTK [25]). This tool is able to align a ground truth reference with a recognized sequence taking into account the timing of the events. This was shown to be very important here due to HBB sounds sometimes being repeated several times in sequence. As scoring measure, we used the error rate, which corresponds to the total number of insertions, deletions and substitutions, divided by the total number of musical events.

4.2. Results

Figures 1a and 1b display the error rate evolution in the proposed range of analysis frame shifts and of HMM state numbers, for individual drum sounds and instrument imitations respectively. Note that the error rate can exceed 100% as it includes insertions, deletions and substitutions. We clearly see, on both figures, that the best performance is achieved using a 2 ms analysis frame shift and a HMM with 21 states, corresponding to error rates of 9.3% and 41% for individual drum sounds and instrument imitations respectively. This contrasts with the typical values in speech processing (frame shift around 5 ms and HMM with 5 states per phoneme in context), and is probably due to the short and more dynamic nature of the signals.

Selecting optimal values from above, HMMs were trained using the proposed range of feature extraction approaches and number of coefficients. Figures 1c and 1d exhibit the error rate evolution when both the coefficient number and the coefficient type vary, for individual drum sounds and instrument imitations respectively.

In the case of individual drum sounds, the best performance is achieved when the model uses 22 MFCCs, leading to a recognition error rate of 9%. This score is detailed in Table 4: recognition accuracy (Corr.), error (Err.), substitution (Sub.), deletion (Del.) and insertion (Ins.) percentages and number of occurrence (Num.).

Table 4: *Detailed results for the automatic recognition of HBB drums and instruments. Optimal setup: frame shift = 2 ms, state # = 21, coefficients = MFCCs and # = 22 (drums) or 18 (instruments).*

Type	Num.	Corr.	Sub.	Del.	Ins.	Err.
Drums	1735	93%	7%	0%	2.1%	9%
Instruments	1381	65.7%	6.6%	27.7%	6.6%	41%

Regarding instrument imitations, the best performance is achieved when the model uses 18 MFCCs, leading to a dramatic drop in performance with a recognition error rate of 41% (Table 4). This can be explained by: i) elec. guit. egr., which is almost never recognized as such (recognition error of 96.3%) and always confused with elec. guit. ing. and guit. bass, and ii) harmonica (recognition error of 47.4%), mostly confused with elec. guit. ing. and trump. cork. Another problem explaining this degradation concerns insertions and deletions. Indeed, as already mentioned earlier, HBB sounds are sometimes repeated several times in sequence, especially for instrument imitations where repetitions only differ by their pitch. As no pitch information is given to the HMM, the task of recognizing individual instrument notes is problematic, and thus, can lead to frequent insertions and/or deletions. Indeed, the percentage of correctly recognized classes for each instrument varies as: 86.6% (trump. cork.), 89.1% (trump. trill.), 92.9% (trump.), 94.6% (didgeridoo), 97.1% (elec. guit. ing.), 98.1% (guit. bass) and even 100% (saxo.).

5. CONCLUSIONS AND FUTURE WORKS

This paper focused on the analysis and automatic recognition of HBB sounds, a challenging task related to spoken language processing. First, a new high-quality HBB audio database was recorded, carefully segmented and annotated manually. We then investigated three complementary aspects of HBB analysis: pitch tracking, onset detection, and automatic recognition of sounds and instruments. First, various state-of-the-art pitch tracking techniques were compared. Applying a majority voting on those algorithm outputs led to interesting results on pitched instruments (GPE of 10.3%, VDE of 9.5% and FFE of 13.3%). Different onset detection methods were also compared. Spectral flux computed on log-magnitude spectra appears as a good compromise (with F-measures reaching 92% and 79% for drums and pitched instruments respectively). HBB sound processing hence appears to be no harder than real instrument sound processing. Finally, HMMs were evaluated for automatic transcription of HBB performance. The HMM parameters space was explored and an optimum was found (in particular, the analysis frame shift of 2 ms), differing strongly from the typical values observed in spoken language processing. Recognition error rates at the musical event level reached 9% and 41% for drums (5 classes) and instrument imitations (8 classes) respectively.

Several future works are planned, e.g.: investigate real-time onset detection and recognition of HBB performances and improving robustness to apply these technologies in less than ideal conditions (practicing outside a soundproof room), and to more complex material (freestyle performances combining multiple drum and instrument sounds). Multimodal aspects will also be considered.

6. ACKNOWLEDGEMENTS

B. Picart is supported by EU, i-Treasures FP7 project (FP7-ICT-2011-9, #600676), and S. Brognaux by FNRS. Authors are grateful to Davox and Matthieu beatbox artists for their availability.

7. REFERENCES

[1] "Intangible treasures - capturing the intangible cultural heritage and learning the rare know-how of living human treasures," http://i-treasures.eu/.

[2] K. Dimitropoulos, S. Manitsaris, F. Tsalakanidou, S. Nikolopoulos, B. Denby, S. Al Kork, L. Crevier-Buchman, C. Pillot-Loiseau, S. Dupont, J. Tilmanne, M. Ott, M. Alivizatou, E. Yilmaz, L. Hadjileontiadis, V. Charisis, O. Deroo, A. Manitsaris, I. Kompatsiaris, and N. Grammalidis, "Capturing the intangible: An introduction to the i-treasures project," in *Proc. of the 9th International Conference on Computer Vision Theory and Applications*, Lisbon, Portugal, January 5-8 2014.

[3] Dan Stowell and Mark D. Plumbley, "Characteristics of the beatboxing vocal style," Tech. Rep., Centre for Digital Music, Dep. of Electronic Engineering, Univ. of London, 2008.

[4] Ajay Kapur, Manjinder S. Benning, and George Tzanetakis, "Query-by-beat-boxing: Music retrieval for the dj," in *Proc. of the International Conference on Music Information Retrieval*, Barcelona, Spain, October 10-15 2004, pp. 170–178.

[5] Elliot Sinyor, Cory McKay, Rebecca Fiebrink, Daniel McEnnis, and Ichiro Fujinaga, "Beatbox classification using ace," in *Proc. of the International Conference on Music Information Retrieval*, London, UK, September 11-15 2005, pp. 672–675.

[6] Dan Stowell and Mark D. Plumbley, "Delayed decision-making in real-time beatbox percussion classification," *Journal of New Music Research*, vol. 39, no. 3, pp. 203–213, 2010.

[7] Michael Proctor, Erik Bresch, Dani Byrd, Krishna Nayak, and Shrikanth Narayanan, "Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study," *Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 1043–1054, February 2013.

[8] Tiphaine de Torcy, Agnè Clouet, Claire Pillot-Loiseau, Jacqueline Vaissière, Daniel Brasnu, and Lise Crevier-Buchman, "A video-fiberscopic study of laryngopharyngeal behaviour in the human beatbox," *Logopedics Phoniatrics Vocology*, 2013.

[9] Onur Babacan, Thomas Drugman, Nicolas d'Alessandro, Nathalie Henrich, and Thierry Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *Proc. of the IEEE ICASSP*, Vancouver, Canada, May 26-31 2013, pp. 7815–7819.

[10] D. Talkin, *Speech Coding and Synthesis*, chapter A robust algorithm for pitch tracking RAPT, pp. 495–518, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier, 1995.

[11] Thomas Drugman and Abeer Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. of Interspeech*, Florence, Italy, August 27-31 2011, pp. 1973–1976.

[12] Alain de Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[13] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. of IFA*, Institute of Phonetic Sciences, University of Amsterdam, 1993, pp. 97–110.

[14] W. Chu and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Proc. of the IEEE ICASSP*, Taipei, Taiwan, 2009, pp. 3969–3972.

[15] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, Sept. 2005.

[16] Paul M. Brossier, *Automatic Annotation of Musical Audio for Interactive Applications*, Ph.D. thesis, Centre for Digital Music Queen Mary, University of London, August 2006.

[17] Paul Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signal*, Ph.D. thesis, University of Bristol, UK, 1996.

[18] Wan-Chi Lee, Yu Shiu, and C.-C. Jay Kuo, "Musical onset detection with joint phase and energy features," in *IEEE International Conference on Multimedia and Expo*, Beijing, China, July 2-5 2007, pp. 184–187.

[19] A. Holzapfel, Y. Stylianou, A. C. Gedik, and B. Bozkurt, "Three dimensions of pitched instrument onset detection," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1517–1527, August 2010.

[20] Simon Dixon, "Onset detection revisited," in *9th International Conference on Digital Audio Effects*, Montreal, Canada, September 28-20 2006, pp. 133–137.

[21] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall Signal Processing Series, 1993.

[22] Mark Gales and Steve Young, "The application of hidden markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, January 2008.

[23] Thierry Dutoit and Stéphane Dupont, "Chapter 3 - speech processing," in *Multimodal Signal Processing*, Jean-Philippe Thiran, Ferran Marqués, and Hervé Bourlard, Eds., pp. 25–61. Academic Press, Oxford, 2010.

[24] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, *The HTK Book (for HTK Version 3.4)*, Cambridge University, 2009.

[25] "Nist scoring toolkit," http://www1.icsi.berkeley.edu/Speech/-docs/sctk-1.2/sctk.htm.