

Session 2pMU

Musical Acoustics and Speech Communication: Musical Pitch Tracking and Sound Source Separation Leading to Automatic Music Transcription II

James W. Beauchamp, Chair
Univ. of Illinois Urbana-Champaign, School of Music, Dept. of Electrical and Computer Engineering, Urbana, IL 61801

Chair's Introduction—1:30

Invited Papers

1:35

2pMU1. Active music listening interfaces based on sound source separation and F0 estimation. Masataka Goto Natl. Inst. of Adv. Industrial Sci. and Technol. AIST, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan, m.goto@aist.go.jp

This paper describes research aimed at building “*active music listening interfaces*” to demonstrate the importance of music understanding technologies, including sound source separation and F0 estimation, and the benefit they offer to end users. Active music listening is a way of listening to music through active interactions. Given polyphonic sound mixtures taken from available music recordings, our active music listening interfaces enrich end-users’ music listening experiences. For example, by suppressing drum sounds and adding other drum sounds, *Drumix* Yoshii et al., *IPSI Journal*, **48** 3, 2007 enables a user to change the volume and timbre of drum sounds separated from the mixtures and rearrange rhythmic patterns of these drum sounds during playback. In addition, during the playback of a song, *LyricSynchronizer* Fujihara et al., *Proc. of IEEE ISM 2006* displays scrolling lyrics and highlights the phrase currently sung. Because lyrics are automatically synchronized with vocal melodies separated from polyphonic sound mixtures using our predominant-F0 estimation method *PreFEST*, a user can easily follow the current playback position and click on a word in the lyrics to listen to it. These interfaces can also be regarded as “*augmented music-understanding interfaces*” that facilitate deeper understanding of music by end users.

1:55

2pMU2. Timbre modeling for the purpose of sound source recognition and separation in music. Anssi Klapuri Inst. of Signal Processing, Tampere Univ. of Technol., Korkeakoulunkatu 1, 33720 Tampere, Finland

Timbre models of musical instruments were studied in order to recognize sound sources in music and to organize sounds to their sources. The log-power spectrum of the sounds was modeled using a combination of linear bases, one defined as a function of frequency, and another as a function of harmonic index. Algorithms for initializing, learning, and sequentially adapting this model are described. In simulations, the model had a benefit over conventional Mel-cepstral representations, especially in the case of certain instruments, including the electric guitar, the organ, the mallet percussions, and the clarinets. The improvement was observed in supervised instrument classification, but not so clearly in the unsupervised clustering of sounds.

2:15

2pMU3. Supervised separation of musical sources. Paris Smaragdis Mitsubishi Electric Res. Labs, 201 Broadway, 8th Fl., Cambridge, MA 02139, paris@media.mit.edu

Sounds, especially in the musical context, are often presented in mixtures. Given that traditional signal processing theory is not well equipped to work with concurrent sound sources, this limits the analysis and creative options on musical audio. As it was recently shown, sound mixtures can be analyzed with latent variable models of time-frequency distributions and, thus, reveal the additive structure of acoustic scenes. Unfortunately, such processes separate acoustic elements in an indiscriminant manner that does not always result in extracting the desired sources. However, in the case where a user can provide examples of the sources to extract, the aforementioned algorithms can become powerful tools for supervised source separation. Performing experiments on real singlechannel recordings, it was shown that by simply indicating the regions of time where a particular source was active, it is possible to extract that source with minimal distortion, even though it may constantly be part of a sound mixture. Since this is an example-based procedure, it is able to deal with arbitrary sources, regardless of their acoustic properties, without requiring any tedious heuristic modeling. This extraction method facilitates the analysis of musical data, and also allows creative manipulation of music.

2:35

2pMU4. A framework for sound source separation using spectral clustering. George Tzanetakis, Mathieu Lagrange, Luis Gustavo Martins, and Jennifer Murdoch Dept. of Comput. Sci., Univ. of Victoria, Canada, gtzan@cs.uvic.ca

Clustering based on the normalized cut criterion, and more generally, spectral clustering methods, are techniques originally proposed to model perceptual grouping tasks, such as image segmentation in computer vision. In this work, it is shown how such techniques can be applied to the problem of dominant melodic source separation in polyphonic music audio signals. One of the main advantages of this approach is the ability to incorporate multiple perceptually-inspired grouping criteria into a single framework without requiring multiple processing stages, as many existing computational auditory science analysis approaches do. Experimental results for several tasks, including dominant melody pitch detection, are presented. The system is based on a sinusoidal modeling analysis front-end. A novel similarity cue based on harmonicity harmonically-wrapped peak similarity is also introduced. The proposed system is data-driven i.e., requires no prior knowledge about the extracted source, causal, robust, practical, and efficient close to real-time on a fast computer. Although a specific implementation is presented, one of the main advantages of the proposed approach is its ability to utilize different analysis front-ends and grouping criteria in a straightforward manner.

2:55

2pMU5. Methods for stereo music source separation. Andreas Ehmann Dept. of Elec. and Comput. Eng., Univ. of Illinois at Urbana-Champaign, 1406 W. Green St., Urbana, IL 61801

Musical source separation from stereo signals has been growing in popularity recently. Algorithms including the author's and the results of a recent stereo audio source separation evaluation campaign SASSEC will be presented. In general, such approaches involve extracting mixing parameters from time-frequency T-F analyses of both channels. Each corresponding T-F point of the two channels is compared and interchannel intensity differences IID and in the case of stereo mixing, interchannel phase differences IPD are computed. A histogram of signal power in the IID and IPD space is constructed, with sources showing strong peaks at the locations of their mixing parameters. Masks are then created for each potential source, corresponding to T-F points that share common mixing parameters. These masks are applied to the T-F analysis and resynthesized as probable sources. The underlying assumption of such systems is that of W-disjoint orthogonality, that is, T-F points of the sources do not overlap. In music, however, this is rarely the case, as common musical intervals often contain overlapping harmonic "collisions." The use of an exponentially damped sinusoidal model, that in limited cases has the ability to resolve very closely spaced, beating, sinusoids is explored as a means of separating harmonic collisions.

3:15–3:25 Break

3:25

2pMU6. Algorithm for separating vocals from polyphonic music. Tuomas Virtanen and Matti Ryynänen Tampere Univ. of Technol., P.O. Box 553, FI-33101 Tampere, Finland, tuomas.virtanen@tut.fi

An algorithm for the separation of vocals from polyphonic music is described. The algorithm consists of two stages, which first estimate the predominant melody line, and then the sinusoidal modeling parameters corresponding to the melody line. The melody line is estimated using a hidden Markov model where the output of a multiple fundamental frequency estimator is used as a feature set. The states of the hidden Markov model correspond to musical notes having different fundamental frequencies, and the state transition probabilities are determined by a musicological model. The sinusoidal modeling stage estimates the frequency, amplitude, and phase of each overtone of the predominant melody line in each frame. The sinusoidal modeling stage can also include mechanisms which reduce the effect of interfering sound sources on the estimated parameters. The resulting separation algorithm is independent of singer identity, musical genre, or instrumentation. Simulation experiments on real polyphonic music show that the algorithm enables separation of vocals from the accompaniment, providing robust results on various musical genres.

3:45

2pMU7. Separation of singing voice from music accompaniment for monaural recordings. Yipeng Li and DeLiang Wang Dept. of Comput. Sci. and Eng., The Ohio State Univ., Columbus, OH 43210

Separating singing voice from music accompaniment is useful in a wide range of applications, such as lyrics recognition and alignment, singer identification, and music information retrieval. Compared to speech separation, which has been extensively studied for decades, singing voice separation has been little explored. We have developed a pitch-based system to separate singing voice from music accompaniment for monaural recordings. Our system consists of three stages. The singing voice detection stage partitions and classifies an input into vocal and nonvocal portions. For each vocal portion, the predominant pitch detection stage detects the pitches of the singing voice. Finally, the separation stage uses the detected pitch to group the time-frequency segments of the singing voice. Quantitative results show that the system performs the separation task successfully. [The work is supported by AFOSR and AFRL.]

4:05

2pMU8. Statistical models for music signal analysis and transcription. A. Taylan Cemgil Dept. of Eng., Univ. of Cambridge, Trumpington St., CB2 1PZ Cambridge, UK

In recent years, there has been an increasing interest in statistical approaches and tools from machine learning, for the analysis of audio and music signals. The application of statistical techniques is quite natural: Acoustical time series can be conveniently modeled using hierarchical signal models by incorporating prior knowledge from various sources: From physics or studies of human cognition and perception. Once a realistic hierarchical model is constructed, many audio processing tasks, such as coding, restoration, transcription, separation, identification, or resynthesis can be formulated consistently as Bayesian posterior inference problems. This contribution illustrates various realistic generative signal models for audio and music signal analysis. In particular, factorial switching state space models, Gamma-Markov random fields, and point process models will be discussed. Some models admit exact inference, otherwise, efficient algorithms based on variational or stochastic approximation methods can be developed. We will illustrate the approach on music transcription, restoration, and source separation applications. [Work supported by EPSRC.]

4:25

2pMU9. Approaching polyphonic transcription of piano sounds. Luis I. Ortiz-Berenguer Audiovisual and Commun. Eng. Dept., EUITT, Polytechnic Univ. of Madrid UPM, Ctra Valencia km7, Madrid, Spain and Francisco J. Casajus-Quiros Polytechnic Univ. of Madrid UPM, Ciudad Universitaria s/n. Madrid, Spain

Polyphonic transcription of piano is a challenging task due to the specific characteristics of its sound spectrum. Pattern matching method, which compares the spectrum with a set of spectral patterns, has proven to give good results, although some limitations still exist mainly when analyzing notes of the lower octaves. Present research is oriented to improve both the accuracy of the spectral patterns, especially for lower octaves, and the matching-score calculation. The former is being carried out by developing a model of the vibration of piano. The last is aimed to decide whether a simple internal product or a more complex MSE calculation are needed and how the spectral pattern has to be modeled regarding partial levels. The model of the vibration permits calculating the partial frequencies, for which it takes into account the vibration of the stiff string and the effect of the soundboard impedance. The soundboard impedance modeling is under a deeper study using both real measurements and FEM modeling. As further work, once the model is completed, it is expected to generate the patterns by using a waveguide synthesis method. [Work supported by Spanish National Project TEC2006-13067-C03-01/TCM.]

Contributed Paper

4:45

2pMU10. Employing the mutual phase coherence of the overtones in musical sounds for audio source separation and reconstruction.
Gordana Velikic and Mark F. Bocko Dept. of Elec. and Comput. Eng., Univ. of Rochester, Rochester, NY 14627

Source separation from a monaural mixture of musical sounds is challenging. When only a single mixture of multiple sources is available, the method of independent component analysis may not be straightforwardly applied, and proposed source separation methods have relied on previous knowledge of the independent sources. Our method employs mutual correlations of the overtone phases of musical sounds to ascribe spectral features to individual sources. The phase versus time is computed from the analytical signal representation of the output of a set of adaptive band-pass filters that perform initial spectral separation. A small number of cleanly separable partials, usually the fundamentals, are then employed as references to which the remaining filter-bank outputs are correlated. To separate overlapping spectral features and ascribe the proper amount of power to each source, the phase of the reference signal is scaled to the harmonic frequency, and the amplitude of each overtone is estimated by computing the cross-correlation of the phase trajectories. To group sonic events in time, we employ the history of the signal itself, thus eliminating the need for prior information. We discuss this separation method and the audio quality of the reconstructed source signals, with examples and a quantitative perceptual metric.