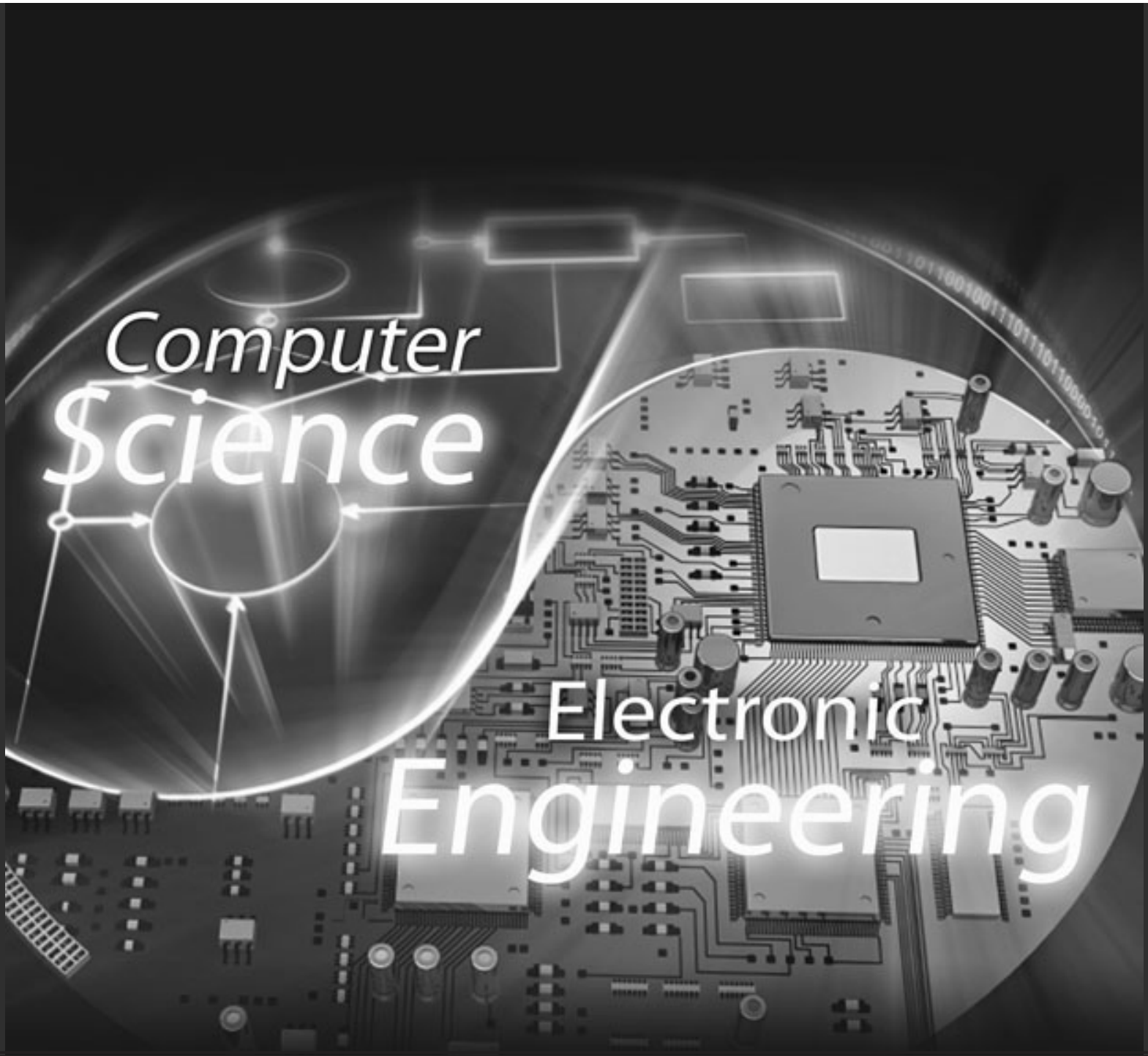
ISSN 2043-0167

Recognition of Harmonic Sounds in Polyphonic Audio using a Missing Feature Approach: Extended Report

Dimitrios Giannoulis, Anssi Klapuri, and Mark D. Plumbley

EECSRR-12-05 November 2012

Queen Mary University of London
School of Electronic Engineering and Computer Science



RECOGNITION OF HARMONIC SOUNDS IN POLYPHONIC AUDIO USING A MISSING FEATURE APPROACH: EXTENDED REPORT

Dimitrios Giannoulis[†]*, Anssi Klapuri*[‡]*, and Mark D. Plumbley*[†]

† Queen Mary University of London, Centre for Digital Music, London, UK
‡ Ovelin, Vilhonkatu 5A, Helsinki, Finland

ABSTRACT

A method based on local spectral features and missing feature techniques is proposed for recognition of harmonic sounds in mixture signals. A mask estimation algorithm is proposed for identifying spectral regions that contain reliable information for each sound source and then bounded marginalization is employed to treat the feature vector elements that are determined as unreliable. The proposed method is tested on musical instrument sounds due to the extensive availability of data but it can be applied on any harmonic sound (i.e. animal sounds, environmental sounds). In simulations the proposed method clearly outperformed a baseline method for mixture signals.

1. INTRODUCTION

Computational auditory scene analysis (CASA) broadly speaking refers to algorithms that aim to recognize sound sources or events in auditory scenes [1]. Applications of CASA include for example intelligent hearing aids, acoustic surveillance, and mobile devices that adapt to the situational context.

In the case of a generic acoustic scene with various types of audio events no system at present has results anywhere close to the results a human listener can achieve as these were measured in early studies [2, 3]. Existing approaches are based on low-level signal features and k-means clustering [4], Hidden Markov Models (HMMS) [5], Probabilistic Latent Semantic Analysis (PLSA) [6], Non-Negative Matrix factorization (NMF) with time-varying bases [7], NMF with time-frequency activations [8], Shift-Invariant Probabilistic Latent Component Analysis (SIPLCA) temporally-constrained via on/off HMMs [9] or local time-frequency patterns and AdaBoost [10]. Many approaches and techniques have been tailored to specific scenes or types of audio signals such as music and speech and the resulting performance of such systems is better and closely comparable to that of humans although there is still room for improvement [11, 12].

The performance of CASA algorithms is significantly affected by the fact that they have to deal with low signal-to-noise ratios and mixtures of multiple overlapping sources. The systems fail where human listeners succeed perhaps because they are unable to imitate the ability of the auditory system to ignore spectrotemporal regions that are corrupted by noise or interfering sources, provided that there is a sufficient amount of information in other regions to suggest the presence of a sound source [13] [14].

Missing feature approaches provide a general framework for recognizing sound sources based on partial information [15, 14, 16]. These techniques attempt to identify spectrotemporal regions that carry reliable information about a sound source, in contrast to regions that are corrupted by interference from other sources or noise and are therefore labeled as unreliable or "missing" [14]. Natural sounds tend to be concentrated in small regions (sparse) in the time-frequency domain and therefore parts of their spectrogram data is often uncorrupted even in the presence of multiple sources.

Missing feature techniques have been applied by a number of authors in environmentally-robust speech recognition (see [15, 16] for reviews), but there has been very little work outside that application domain. Arguably one of the main reasons for that is the difficulty of estimating the "mask" that identifies reliable and unreliable (noisy) spectrotemporal regions: in CASA hardly any assumptions can be made about the target sounds or the interference (in contrast to environmentally-robust speech recognition). In musical instrument recognition, Eggink and Brown employed a missing feature approach by using pitch information to predict harmonic partial collisions and thereby estimate the mask [17].

In the following, we propose a missing feature algorithm for recognizing harmonic sounds in mixture signals. As acoustic features in the proposed method, we use log-energy differences between spectral subbands. For mask estimation, we use a novel technique based on spectral smoothness. Unreliable feature vector elements are handled using bounded marginalization. The recognition is currently performed independently in each frame but it is possible to extend the method to include temporal features and to integrate information over time. In mixture signals, the proposed method clearly outperforms a reference Bayesian classifier based on Mel-cepstral features.

This report serves as an extended version for the ICASSP submission [18].

2. METHOD

Let us denote the observed audio signal at time frame t by vector $\mathbf{o}_t = [o_t(n)]_{n=1,\ldots,N}$. The observation is modelled as a mixture of harmonic sounds and a residual:

$$\mathbf{o}_t = \sum_{f \in \mathcal{F}_t} \mathbf{s}_{f,t} + \mathbf{r}_t \tag{1}$$

where f denotes the pitch of sound $\mathbf{s}_{f,t} = [s_{f,t}(n)]_{n=1,\ldots,N}$, and the set $\mathcal{F}_t$ contains all pitches of sounds that are active in frame t. The residual signal $\mathbf{r}_t$ represents all non-harmonic sounds such as background noise. For convenience, we omit the frame index t in the following and write (1) as $\mathbf{o} = \sum_{f \in \mathcal{F}} \mathbf{s}_f + \mathbf{r}$.

Let us use $\mathsf{p}(c|\mathbf{o})$ to denote the probability that class c is present in frame $\mathbf{o}$ and $\mathsf{p}(c_f\mathbf{o})$ to denote the probability that sound $\mathbf{s}_f$ belongs to class c. Using the complements $\mathsf{p}(c|\mathbf{o}) = 1 - \bar{\mathsf{p}}(c|\mathbf{o})$ and $\mathsf{p}(c_f|\mathbf{o}) = 1 - \bar{\mathsf{p}}(c_f|\mathbf{o})$ and assuming all sounds are independent we can write

$$\bar{\mathsf{p}}(c|\mathbf{o}) = \prod_{f \in \mathcal{F}} \bar{\mathsf{p}}(c_f|\mathbf{o}) \tag{2}$$

The above equation says that one sound from class c suffices to conclude that class c is present. We are further making the assumption that the set of active pitches $\mathcal{F}$ is estimated reliably, that the set $\mathcal{F}$ is "given." We then model $\mathsf{p}(c_f|\mathbf{o})$ by

$$\mathsf{p}(c_f|\mathbf{o}) = \mathsf{p}(c_f|\mathbf{y}_f, f) \tag{3}$$

where the observation $\mathbf{o}$ has been replaced by feature vector $\mathbf{y}_f$ that is extracted from the mixture signal to represent the sound with pitch f based on the assumption that $\mathbf{y}_f$ sufficiently describess all the information of the sound $\mathbf{s}_f$. The probabilities $\mathsf{p}(f|\mathbf{o})$ of different pitches f to be present are obtained using a multipitch estimation method such as the one described in [19]. For simplicity, as explained above, we assume that the probabilities $\mathsf{p}(f|\mathbf{o})$ of those pitches are 1, that is, that we are certain those are the set of active pitches in the frame, that is, the set $\mathcal{F}$ is "given"

The focus of this paper is on calculating $\mathsf{p}(c_f|\mathbf{y}_f, f)$, that is, the probability that a candidate sound belongs to class c when given its pitch f and the feature vector $\mathbf{y}_f$ extracetd from the mixture signal $\mathbf{o}$ (as will be explained below). The problem becomes non-trivial and thus interesting in polyphonic scenarios where the feature vector $\mathbf{y}_f$ is usually partly obscured by other co-occurring sounds that overlap in the time-frequency domain. Probabilistic models representing instrument c are trained using *clean* feature vectors extracted from isolated signals representing instrument c. This is because the interference caused by other, co-occurring sounds in polyphonic audio is highly varying and unpredictable and therefore any interference introduced at the training stage would hardly be representative of the test stage.

The problem can then be re-stated as calculating the probability $\mathsf{p}(c_f|\mathbf{y}_f, f)$ when some elements of the feature vector $\mathbf{y}_f$ are reliable (clean) and some are obscured. The reliability information ("mask") is generally not available for mixture signals but has to be estimated too.

2.1. Feature representation and Binary mask

A variety of acoustic features have been proposed for audio classification, including spectral, cepstral, temporal, and modulation-spectral features [20]. In the missing feature framework, the features should be local in time-frequency in order to be able to avoid interfering sounds that tend to have a sparse energy distribution in that domain and therefore only a local effect on the features. We use log-energy differences between spectral subbands, which removes the need for level normalization and ensures that interference remains local to specific spectrum areas as opposed to cepstral features where it would spread all over the feature vector. The feature vector $\mathbf{y}_f$ is calculated by first picking the harmonic partials of a sound with pitch f from the mixture spectrum by assuming that the frequencies of the partials are exact integer multiples of the estimated pitch. We have found that extracting spectral energy only at the positions of the partials considerably improves the signal-to-noise ratio from the viewpoint of the candidate sound with pitch f.

Let vector $\mathbf{x}_f = [x_f(h)]_{h=1,2,\ldots,H}$ denote the powers of harmonic partials h in the observed mixture spectrum at frequencies hf. The actual feature vector is then obtained from

$$\mathbf{y}_f = 10\log_{10}(\mathbf{B}\mathbf{x}_f) \tag{4}$$

where the transform matrix $\mathbf{B}$ maps from a linear to log-frequency resolution in order to reduce the dimensionality and also improves the statistical properties of the features. The matrix is given by

$$[\mathbf{B}]_{k,h} = g\left(\frac{\pi}{\gamma}\log_2\left(\frac{\omega_k}{hf}\right)\right) \tag{5}$$

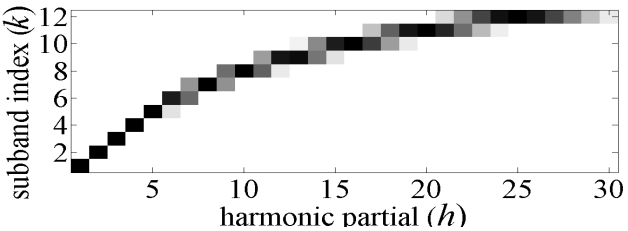


Fig. 1. Illustration of the transformation matrix $[\mathbf{B}]_{k,h}$ for $\gamma = \frac{1}{3}$.

where the window function $g(a) = 0.5 + 0.5\cos(a)$ for $a \in [-\pi, \pi]$ and zero elsewhere. k denotes the elements of the feature vector $\mathbf{y}_f$ refered to as *subbands* in the following. Parameter γ determines the log-frequency resolution of the features, for example $\gamma = \frac{1}{3}$ leads to a third-octave resolution. For small γ, $\mathbf{B}$ becomes an identity matrix.

Center frequencies ω_k of the subbands depend on pitch f are defined recursively by setting $\omega_1 = f$ and $\omega_k = \max(2^\gamma\omega_{k-1}, \omega_{k-1} + f)$. This ensures that all elements of $y_f(k)$ are informative. Figure 1 illustrates matrix $\mathbf{B}$ for $\gamma = \frac{1}{3}$.

The feature vector $\mathbf{y}_f$ extracted from the mixture signal $\mathbf{o}$ is likely to be partly obscured by other sounds that overlap the target sounds in time and frequency.

Let us use $\mathbf{z}_f$ to denote the unobserved, "clean," feature vector that we would obtain if the features were extracted from sound $\mathbf{s}_f$ in isolation. Let us define binary masks $m_f(k)$, where $m_f(k) = 1$ indicates that the measured log-power $y_f(k)$ for subband k is dominated by energy coming from the source with pitch f. More exactly, we assume that the (unobserved) clean feature vector $\mathbf{z}_f$ obeys

$$\begin{aligned} z_f(k) &= y_f(k) \;\text{ if } m_f(k) = 1 \\ z_f(k) &\le y_f(k) \;\text{ if } m_f(k) = 0 \end{aligned} \tag{6}$$

The latter stems from the fact that the *expected value* of the power spectrum of the mixture signal $\mathbf{o}$ is the sum of the power spectra of sources $\mathbf{s}_f$, $f \in \mathcal{F}$. This is valid only in the expectation sense, but is a useful assumption for classification purposes as will be seen.

Estimating the masks $\mathbf{m}_f$ of each sound from the observed mixture signal will be discussed in Sec. 3.1. The clean "glimpses" of the sources, when $m_f(k) = 1$, form a basis for the recognition. However also the subbands where $m_f(k) = 0$ inform about $\mathbf{s}_f$: the observed feature value $y_f(k)$ sets an upper bound for the unobserved clean feature value $z_f(k)$. To keep the notation uncluttered, we omit the subscript f in the following and write simply $\mathbf{z}$, $\mathbf{y}$, and $\mathbf{m}$, with the exception of c_f to avoid confusion with c.

2.2. Marginalization of the missing data

The marginalization approach explained in this subsection is similar to the one proposed in [21], although the employed model and features are different. The probability $\mathsf{p}(c_f|\mathbf{y}, f)$ that a candidate sound $\mathbf{s}_f$ belongs to class c, as required in (3), can be written as $\mathsf{p}(c_f|\mathbf{y}, f) = \sum_{\mathbf{m}} \mathsf{p}(c_f|\mathbf{m}, \mathbf{y}, f)p(\mathbf{m})$. In the case of a deterministic mask estimation we can set its probability $p(\mathbf{m}) = 1$, leading to $\mathsf{p}(c_f|\mathbf{y}, f) = \mathsf{p}(c_f|\mathbf{m}, \mathbf{y}, f)$. That can be written as

$$\begin{aligned} \mathsf{p}(c_f|\mathbf{m}, \mathbf{y}, f) &= \int \mathsf{p}(c_f, \mathbf{z}|\mathbf{m}, \mathbf{y}, f)\mathrm{d}\mathbf{z} \\ &= \int \mathsf{p}(c_f|\mathbf{z}, \mathbf{m}, \mathbf{y}, f)\mathsf{p}(\mathbf{z}|\mathbf{m}, \mathbf{y}, f)\mathrm{d}\mathbf{z} \end{aligned} \tag{7}$$

where $\mathsf{p}(\mathbf{z}|\mathbf{m}, \mathbf{y}, f)$ is given by (9) and the integral is used to marginalize $\mathbf{z}$. The factor $\mathsf{p}(c_f|\mathbf{z}, \mathbf{m}, \mathbf{y}, f)$ simplifies to $\mathsf{p}(c_f|\mathbf{z}, f)$ since c_f does not depend on $\mathbf{m}$ or $\mathbf{y}$ given $\mathbf{z}$.

Using Bayes' rule for $\mathsf{p}(c_f|\mathbf{z}, f)$, (7) becomes

$$\mathsf{p}(c_f|\mathbf{m}, \mathbf{y}, f) = \mathsf{p}(c_f|f) \int \frac{\mathsf{p}(\mathbf{z}|c_f, f)}{\mathsf{p}(\mathbf{z}|f)} \mathsf{p}(\mathbf{z}|\mathbf{m}, \mathbf{y}, f) \mathrm{d}\mathbf{z} \quad (8)$$

where $\mathsf{p}(c_f|f)$ is the prior probability of sound with class c at pitch f and $\mathsf{p}(\mathbf{z}|c_f, f)$ is the likelihood of observing $\mathbf{z}$ for sound of class c and pitch f. The latter can be estimated from training data representing *isolated* (clean) signals from class c. The pdf $\mathsf{p}(\mathbf{z}|f)$ is estimated similarly but using data from all classes.

The assumptions in (6) allow us to write the probability density function (pdf) of the unobserved clean features $\mathbf{z}$ of sound $\mathbf{s}_f$:

$$\mathsf{p}(z(k)|\mathbf{m}, \mathbf{y}, f) \quad (9)$$
$$= \begin{cases} \delta(z(k) - y(k)) & \text{if } m(k) = 1 \\ U\mathsf{p}(z(k)|\boldsymbol{\mu}, \boldsymbol{\upsilon}, f) & \text{if } m(k) = 0 \text{ and } z(k) \leq y(k) \\ 0 & \text{if } m(k) = 0 \text{ and } z(k) > y(k) \end{cases}$$

where $\delta(\cdot)$ is the Dirac delta function and U is a normalizing constant to make the pdf sum to unity since the pdf is truncated to be zero above $y(k)$. $\mathsf{p}(z(k)|\boldsymbol{\mu}, \boldsymbol{\upsilon}, f)$ is the distribution of $z(k)$ given values of $\mathbf{z}$ at subbands where $m(k) = 1$. $\boldsymbol{\mu}$ denotes a tuple of subband indices k ordered from smallest to largest, where $m(k) = 1$, so that $m(k) = 1$ if and only if $k \in \boldsymbol{\mu}$. The corresponding values of $\mathbf{y}$ are stored in set $\boldsymbol{\upsilon}$ so that $\upsilon(k) = y(\mu(k))$. The distribution $\mathsf{p}(z(k)|\boldsymbol{\mu}, \boldsymbol{\upsilon}, f)$ is learned using isolated sounds from all different classes.

The above-described approach for computing $\mathsf{p}(c_f|\mathbf{y}, f)$ is theoretically satisfying but requires two problems to be solved in order to be practically useful. Firstly, the statistical models for $\mathsf{p}(\mathbf{z}|\mathbf{m}, \mathbf{y}, f)$ should be invariant to the presentation level (scaling) of sound $\mathbf{s}_f$, appearing as an additive constant in the log-power features $\mathbf{z}$. (Note that we cannot normalize the scale since some of the feature vector elements are obscured and therefore not available.) To achieve that, we consider only level *differences* between subbands k. Let us use $\mathrm{d}_k^\ell \equiv z(k) - z(\ell)$ as a shorthand to denote the level difference between subbands k and ℓ.

Secondly, the multi-dimensional integral over $\mathbf{z}$ in (8) is not computationally feasible in a direct form. The following addresses the computational complexity of the integral in (8) by deriving a factorial form instead.

3. FACTORIAL FORM FOR THE MULTIDIMENSIONAL DENSITY

The factor in the brackets in (8) is assumed to take the following factorial form

$$\int \frac{\mathsf{p}(\mathbf{z}|c_f, f)}{\mathsf{p}(\mathbf{z}|f)} \mathsf{p}(\mathbf{z}|\mathbf{m}, \mathbf{y}, f) \mathrm{d}\mathbf{z} \quad (10)$$
$$= \int \frac{\mathsf{p}(\mathbf{z}_{\boldsymbol{\mu}}|c_f, f)}{\mathsf{p}(\mathbf{z}_{\boldsymbol{\mu}}|f)} \mathsf{p}(\mathbf{z}_{\boldsymbol{\mu}}|\mathbf{m}, \mathbf{y}, f)$$
$$\times \prod_{k \notin \boldsymbol{\mu}} \frac{\mathsf{p}(z(k)|\mathbf{z}_{\boldsymbol{\mu}}, c_f, f)}{\mathsf{p}(z(k)|\mathbf{z}_{\boldsymbol{\mu}}, f)} \mathsf{p}(z(k)|\mathbf{m}, \mathbf{y}, f) \mathrm{d}\mathbf{z}$$

where we have used $\mathbf{z}_{\boldsymbol{\mu}}$ to denote a shorter feature vector containing only the elements at clean subbands, $k \in \boldsymbol{\mu}$. For the other elements, the above equation assumes that $z(k)$ for $k \notin \boldsymbol{\mu}$ are independent of each other given f and the values at the clean subbands.

The above assumption allows us to put the integral inside the product in (10), writing it as

$$\underbrace{\frac{\mathsf{p}(\boldsymbol{\upsilon}|c_f, f)}{\mathsf{p}(\boldsymbol{\upsilon}|f)}}_{\text{clean subbands}} \underbrace{\prod_{k \notin \boldsymbol{\mu}} \int \frac{\mathsf{p}(z(k)|\boldsymbol{\upsilon}, c_f, f)}{\mathsf{p}(z(k)|\boldsymbol{\upsilon}, f)} \mathsf{p}(z(k)|\mathbf{m}, \mathbf{y}, f) \mathrm{d}z(k)}_{\text{noisy subbands}} \quad (11)$$

where for the clean bands we have used $\mathsf{p}(\mathbf{z}_{\boldsymbol{\mu}}|\mathbf{m}, \mathbf{y}, f) = \delta(\mathbf{z}_{\boldsymbol{\mu}} - \boldsymbol{\upsilon})$ from the probability density function (pdf) of the unobserved clean features $\mathbf{z}$ of sound $\mathbf{s}_f$ in (9).

The integral in (11) is over each element of $z(k)$ separately and this makes the (originally multidimensional) integral tractable.

Another important requirement is that $\mathsf{p}(\mathbf{z}|\mathbf{m}, \mathbf{y}, f)$ should be invariant to the presentation level (scaling) of sound $\mathbf{s}_f$, appearing as an additive constant in the log-power features $\mathbf{z}$. (Note that we cannot normalize the scale since some of feature vector elements are obscured and therefore not available.) To achieve that, we only consider level *differences* between subbands k. Let us use $\mathrm{d}_k^\ell \equiv z(k) - z(\ell)$ as a shorthand to denote the level difference between subbands k and ℓ.

We assume that the level difference $\mathrm{d}_{\mu(i)}^{\mu(i+1)}$ of each neighbouring pair of clean subbands depends only on f and the level differences on both sides, $\mathrm{d}_{\mu(i-1)}^{\mu(i)}$ and $\mathrm{d}_{\mu(i+1)}^{\mu(i+2)}$, but not on other subbands. This assumption is made for computational tractability. Retaining the dependency on the differences on both sides is important since the mentioned differences share one subband and are therefore strongly correlated. We can then write the part indicated as "clean subbands" in (11) as

$$\frac{\mathsf{p}(\boldsymbol{\upsilon}|c_f, f)}{\mathsf{p}(\boldsymbol{\upsilon}|f)} = \frac{\mathsf{p}\left(\mathrm{d}_{\mu(1)}^{\mu(2)}|c_f, f\right)}{\mathsf{p}\left(\mathrm{d}_{\mu(1)}^{\mu(2)}|f\right)} \prod_{i=2}^{|\boldsymbol{\mu}|-1} \frac{\mathsf{p}\left(\mathrm{d}_{\mu(i)}^{\mu(i+1)}|\mathrm{d}_{\mu(i-1)}^{\mu(i)}, c_f, f\right)}{\mathsf{p}\left(\mathrm{d}_{\mu(i)}^{\mu(i+1)}|\mathrm{d}_{\mu(i-1)}^{\mu(i)}, f\right)}$$
$$\stackrel{\text{def.}}{=} P_{\mu(1),\mu(2)}^{c_f,f} \prod_{i=2}^{|\boldsymbol{\mu}|-1} P_{\mu(i),\mu(i+1)}^{c_f,f} \quad (12)$$

where on the last row we have introduced the shorthand notation $P_{k,\ell}^{c_f,f}$ for convenience in the following. In the special case where all subbands are clean (i.e., the mask is all-one), (11) would reduce to $P_{1,2}^{c_f,f} \prod_{i=2}^{K-1} P_{i,i+1}^{c_f,f}$.

For the noisy bands, $k \notin \boldsymbol{\mu}$, we calculate the level difference $\mathrm{d}_k^{\alpha(k)}$ between band k and its nearest clean subband $\alpha(k)$. More precisely, $\alpha(k) = \arg\min_{\ell \in \boldsymbol{\mu}}(k - \ell)$ denotes member in set $\boldsymbol{\mu}$ that is nearest to k. The subband $\alpha(k)$ is used as a "point of reference" for band k for which $m(k) = 0$. Similarly, $\beta(k)$ is used to denote the second-nearest member of $\boldsymbol{\mu}$ to k.

We assume that $\mathrm{d}_k^{\alpha(k)}$ depends only on f and the level difference $\mathrm{d}_{\alpha(k)}^{\beta(k)}$ between the two nearest clean subbands, but not on the other bands. As a result, the part indicated as "noisy bands" in (11) can be written as

$$\prod_{k \notin \boldsymbol{\mu}} \int \frac{\mathsf{p}\left(\mathrm{d}_k^{\alpha(k)}|\mathrm{d}_{\alpha(k)}^{\beta(k)}, c_f, f\right)}{\mathsf{p}\left(\mathrm{d}_k^{\alpha(k)}|\mathrm{d}_{\alpha(k)}^{\beta(k)}, f\right)} \mathsf{p}(\mathrm{d}_k^{\alpha(k)}|\mathrm{d}_{\alpha(k)}^{\beta(k)}, \mathbf{m}, \mathbf{y}, f) \mathrm{d}z(k)$$
$$\stackrel{\text{def.}}{=} \prod_{k \notin \boldsymbol{\mu}} Q_{k,\alpha(k),\beta(k)}^{c_f,f} \quad (13)$$

Based on (9), we can write the pdf of $\mathrm{d}_k^{\alpha(k)}$ as

$$\mathsf{p}(\mathrm{d}_k^{\alpha(k)}|\mathrm{d}_{\alpha(k)}^{\beta(k)}, \mathbf{m}, \mathbf{y}, f) = \begin{cases} \frac{1}{W}\mathsf{p}\left(\mathrm{d}_k^{\alpha(k)}|\mathrm{d}_{\alpha(k)}^{\beta(k)}, f\right) & \text{for } z(k) \le y(k) \\ 0 & \text{for } z(k) > y(k) \end{cases} \quad (14)$$

where recall that d_k^ℓ is just a shorthand for $z(k) - z(\ell)$. The normalizing constant W is required because the pdf is truncated. Its value is

$$W = \int_{-\infty}^{y(k)} \mathsf{p}\left(\mathrm{d}_k^{\alpha(k)}|\mathrm{d}_{\alpha(k)}^{\beta(k)}, f\right) \quad (15)$$

Substituting (14)–(15) into (13) we get

$$Q_{k,\alpha(k),\beta(k)}^{c_f,f} = \frac{\int_{-\infty}^{y(k)} \mathsf{p}\left(\mathrm{d}_k^{\alpha(k)}|\mathrm{d}_{\alpha(k)}^{\beta(k)}, c_f, f\right)\mathrm{d}z(k)}{\int_{-\infty}^{y(k)} \mathsf{p}\left(\mathrm{d}_k^{\alpha(k)}|\mathrm{d}_{\alpha(k)}^{\beta(k)}, f\right)\mathrm{d}z(k)} \quad (16)$$

Finally, substituting (11)–(13) into (8), we can write $\mathsf{p}(c_f|\mathbf{m}, \mathbf{y}, f)$ as

$$\mathsf{p}(c_f|\mathbf{m}, \mathbf{y}, f) = \underbrace{\mathsf{p}(c_f|f)}_{\text{class prior}} \underbrace{\left[\prod_{i=1}^{|\boldsymbol{\mu}|-1} P_{\mu(i),\mu(i+1)}^{c_f,f}\right]}_{\text{clean subbands}} \underbrace{\left[\prod_{k\notin\boldsymbol{\mu}} Q_{k,\alpha(k),\beta(k)}^{c_f,f}\right]}_{\text{noisy subbands}} \quad (17)$$

Calculation of the terms $P_{k,\ell}^{c_f,f}$ and $Q_{k,\ell,j}^{c_f,f}$ requires estimating the distributions $\mathsf{p}(\mathrm{d}_k^\ell|\mathrm{d}_j^k, c_f, f)$ from training data. In practice, the joint distributions $\mathsf{p}(\mathrm{d}_k^\ell, \mathrm{d}_j^k|c_f, f)$ are estimated for all possible triplets j, k, ℓ, separately for all different classes c, and for the case where the distributions are not conditioned on the class at all, that is, from training material repesenting all classes.

We use a multivariate Gaussian distribution with full (2×2) covariance matrices to model the densities $\mathsf{p}(\mathrm{d}_k^\ell, \mathrm{d}_j^k|c_f, f)$. This renders the conditional distribution $\mathsf{p}(\mathrm{d}_k^\ell|\mathrm{d}_j^k, c_f, f)$ to be univariate Gaussian by doing the following:

Let $\mathbf{x} = [x_1, x_2]^\mathsf{T}$ denote a multivariate normal random variable with mean $\boldsymbol{\mu} = [\mu_1, \mu_2]^\mathsf{T}$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \quad (18)$$

The conditional distribution $\mathsf{p}(x_1|x_2 = a)$ of x_1 given $x_2 = a$ is normally distributed with mean

$$\hat{\mu} = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(a - \mu_2) \quad (19)$$

and variance

$$\hat{\sigma} = \sigma_{11} - \frac{\sigma_{12}\sigma_{21}}{\sigma_{22}}. \quad (20)$$

The value of the integral in (16) is then obtained from the Gaussian cumulative distribution.

3.1. Mask estimation

Mask estimation is a central and arguably the most difficult part of missing feature techniques. A number of methods have been proposed in the field of environmentally robust speech recognition [16, 21, 22, 23]. However these approaches are less straightforward to apply in CASA where the interference is usually not slowly-varying and does not represent a single source but instead both the target and the interference often belong to the same broad class of environmental sounds.

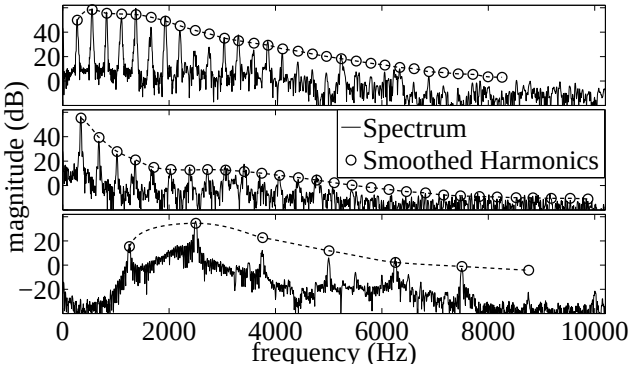


Fig. 2. The spectrum of a musical instrument sound (top), a Humback whale call (middle) and a modern cordless phone ringing (bottom). The smoothed harmonic partial magnitudes $a(h)$ have been highlighted with "$\circ$" and are connected with line segments to produce the "smooth envelope" of the sound.

The mask estimation algorithm proposed in the following is based on the assumption that the spectral envelopes of natural sounds tend to be *smooth*: slowly-varying as a function of log-frequency, in a specific sense [24]. The amplitude of an individual frequency partial can deviate negatively from the smooth envelope, but is very seldom much higher than those of its neighbours. In the latter case, the partial it is more easily perceptually segregated and perceived as a separate sound. This is particularly true for musical instrument sounds, but also for many other natural or artificial sounds. Figure 2 shows examples of smooth spectra for various harmonic sounds.

Overtone partials overlapping with a more dominant partial (from another source) tend to have higher magnitudes than their neighbours and rise above the smooth spectral envelope. That suggests a heuristic that individual partials with amplitudes clearly higher than their neighbours are more likely to have been corrupted by partials from interfering sources, and the mask value at the corresponding position of the feature vector should be set to zero. That is the basic idea of the mask estimation procedure in the following.

The algorithm first estimates the smooth spectral envelope by calculating a local moving average over the series of observed harmonic amplitudes $[x(h)]^{1/2}$ of a sound (recall from Sec. 2.1 that $x(h)$ denotes the power of partial h). An octave-wide Hamming window is centered at each harmonic partial h, and a weighted average $a(h)$ of the partial magnitudes within the window is calculated. The smoothed magnitude spectrum values $a(h)$ are then squared and $[a(h)]^2$ are substituted for $x(h)$ in (4) in order to get a feature vector $\mathbf{y}_{\mathsf{smo}}$. We propose to estimate the mask directly based on the difference $\Delta(k) = y(k) - y_{\mathsf{smo}}(k)$. The mask estimate is given by

$$\hat{m}(k) = \begin{cases} 1 & \text{if } \Delta(k) \le \epsilon_{\mathsf{smo}} \\ 0 & \text{if } \Delta(k) > \epsilon_{\mathsf{smo}} \end{cases} \quad (21)$$

the threshold value $\epsilon_{\mathsf{smo}} = 3$ dB was chosen based on preliminary experiments.

In order to compare the performance of the estimated mask, we utilize an "oracle" mask: an underlying ideal mask. The oracle mask is available at the training stage by generating training sound mixtures for which we have the isolated (clean) sounds before mixing.

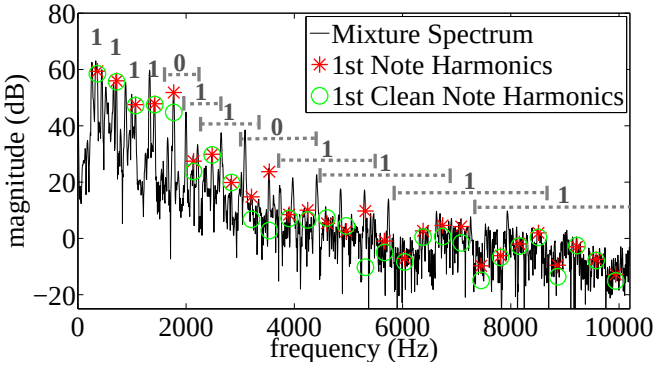


Fig. 3. The spectrum of a mixture of 4 musical instrument sounds. The observed partial magnitudes $y(k)$ of a flute sound of the mixture have been indicated with "$*$" and the underlying clean magnitudes $z(k)$ with "$\circ$". The ideal ("oracle") mask for the sound is indicated with 1s and 0s above the spectrum and the subband boundaries are shown with dotted horizontal lines.

We compute the feature vectors $\mathbf{y}$ from the mixture, and in addition we compute the clean feature vectors $\mathbf{z}$ by applying the same feature extraction procedure on the isolated sounds before mixing. The oracle mask is then defined as

$$m_{\text{or}}(k) = \begin{cases} 1 & \text{if } |y(k) - z(k)| \leq \epsilon \\ 0 & \text{if } |y(k) - z(k)| > \epsilon \end{cases} \quad (22)$$

where ϵ is an empirically found threshold value (in preliminary experiments, values 3–6 [dB] were found suitable).

The top panel of Figure 3 shows an example spectrum consisting of the polyphonic mixture of four instruments. The observed partial magnitudes $y(k)$ of a flute sound have been indicated with "$*$" and the underlying clean magnitudes $z(k)$ with "$\circ$". The ideal (oracle) mask for this sound is shown as a series of 1s and 0s above the spectrum. The subband boundaries are indicated with dotted horizontal lines (note that subbands 1–4 contain a single partial only).

4. SIMULATION RESULTS

For practical purposes (mainly the availability of data), we use musical instrument sounds as the target classes but the method is not limited to musical sounds. Musical instruments provide a wide range of well-defined sound source classes with a lot of acoustic variability within each class. We used the RWC Musical Instrument Sound database [25] for training the class models, and another database, MGill University Master Samples [26] at the test stage. Ten different instruments, available in both, were chosen: bassoon, cello, clarinet, flute, oboe, piano, piccolo, alto saxophone, tuba and violin.

As a baseline method, we employed a Bayesian classifier using Gaussian mixture models (GMMs) to represent the class-conditional likelihood densities (10 Gaussians per model and diagonal covariance matrices). The feature vector was consisted of Mel-frequency cepstral coefficients (MFCCs), which have been widely used for musical instrument recognition [27] and speech recognition [28]. The zeroth coefficient was discarded and the following 12 coefficients were used for classification. The features were element-wise mean and variance normalized over all the training data.

At the test stage, single instrument sounds were mixed with background noise from four different auditory scenes: rain and rumble, crowded bar, dishwashing, and shower. Audio data for these

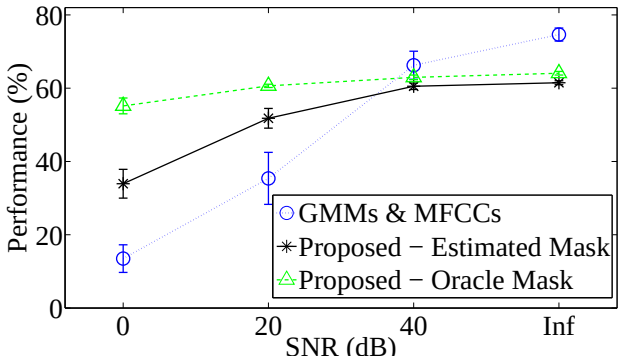


Fig. 4. Performance of the different systems under varied SNR conditions. Mean values and standard deviations out of 12 randomly sampled datasets using the four acoustic scenes.

were obtained from Freesound.org [29]. Recognition was carried out in an individual 93 ms analysis frame. Figure 4 shows results for the proposed method and the baseline method for varying signal-to-noise ratios. To analyze the effect of mask estimation errors, results are also shown for the proposed method with the "oracle" mask, that is obtained by utilizing signal information before mixing as desribed in 3.1. The proposed method outperforms the reference method by a wide margin in low SNR. The full potential of the proposed method can be appreciated by seeing the robustness with the oracle mask.

Table 1 shows results for mixtures of musical instrument sounds without background noise. In this case, the interference is due to the other co-occurring sources. Random notes from random instruments were chosen to generate 10000 one-, two-, and four-sounds mixtures. We had to constrain the test mixtures so that each instrument appears only once in a given mixture. This information, along with the number of sounds in the mixture ("polyphony"), was given as side-information to the classifiers. This was unavoidable since the baseline classifier operates by simply choosing P most probable classes to the output. As a consequence, the random guess rate for isolated sounds is 10% but 40% for four-sound mixtures (guessing 4 out of 10 instruments). The baseline method was trained using mixture signals of the same polyphony as the test material in each case as this led to much better results than training from isolated samples.

The proposed method (last row) outperforms the baseline (row 2) by a wide margin for polyphonies 2 and 4. For clean isolated samples, however, the proposed method performs clearly worse than the GMM+MFCC baseline. The main reason is that the proposed features are based on the amplitudes of harmonic partials only, discarding the spectrum between the partials and also being subject to pitch estimation errors. This conclusion was verified by computing MFCCs using only the harmonic partials of the sound only, setting

Table 1. Recognition accuracy (%) of different methods.

Method		Polyphony		
Model & Features	Mask	1	2	4
1. Random guess	–	10.0	20.0	40.0
2. **GMM & MFCC**	–	74.6	**50.7**	**53.1**
3. GMM	MFCC-H	62.3	46.5	51.8
4. Proposed	Oracle	64.1	62.8	67.5
5. Proposed	Oracle(full m.)	64.1	60.3	64.6
6. Proposed	All-one	64.1	51.8	56.4
7. **Proposed**	**Estimated**	61.5	**56.9**	**60.2**

the spectrum between the partials to zero ("MFCC-H" on row 3).

Rows 4–7 of Table 1 show results for the proposed method. Three different types of masks were tested: the "oracle" mask (row 4), the estimated mask (bottom row), and an all-one mask that assumes all subbands are clean (row 6). The performance difference between the oracle and all-one mask is quite drastic for polyphonies 2 and 4, highlighting the importance of handling unreliable data appropriately in the classification process. Results for the estimated mask are approximately half-way between the oracle mask and the all-one mask, indicating that the spectral smoothness-based mask estimation is able to make an important step towards the ideal mask.

Finally, we calculated results for the oracle mask using the proposed bounded integration (row 4) but also full marginalization (row 5), where the integral over $z(k)$ is calculated from $-\infty$ to ∞ instead of $-\infty$ to $y(k)$, which has the consequence that the noise terms $Q_{j,k,\ell}^{c_f,f}$ become one and need not be computed at all. Bounded marginalization works consistently better than full marginalization.

5. CONCLUSION

In this paper we proposed a novel method for identification of harmonic sounds in polyphonic mixtures. The method is based on the missing feature approach and local spectral features, using bounded marginalization to treat the unreliable feature vector elements. A mask estimation technique was proposed that is based on the assumption that the spectral envelopes of musical sounds tend to be slowly-varying as a function of log-frequency.

The proposed method outperformed the reference method (GMM+MFCC) clearly in mixture signals. For isolated samples, the proposed method performed somewhat worse than the reference method, which seems to be due to the fact that only information at the positions of the harmonic partials is utilized and the rest of the spectrum is discarded.

6. REFERENCES

[1] D.L. Wang and G.J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, IEEE Press, 2006.

[2] V.T.K. Peltonen, A.J. Eronen, M.P. Parviainen, and A.P. Klapuri, "Recognition of everyday auditory scenes: potentials, latencies and cues," *Preprints - Audio Engineering Society (AES)*, 2001.

[3] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[4] A. Harma, M.F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, p. 4.

[5] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *Proc. of 18th European Signal Processing Conference*, 2010.

[6] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection," *Proc. of 19th European Signal Processing Conference (EUSIPCO)*, 2011.

[7] C.V. Cotton and D.P.W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 69–72.

[8] R. Hennequin, R. Badeau, and B. David, "NMF with time–frequency activations to model nonstationary audio events," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744–753, 2011.

[9] E. Benetos, M. Lagrange, and S. Dixon, "Characterisation of acoustic scenes using a temporally-constrained shift-invariant model," *Proc. of the 15 Int. Conference on Digital Audio Effects (DAFx-12)*, September 17-21 2012.

[10] A. Härmä, "Detection of audio events by boosted learning of local time-frequency patterns," *Watermark*, vol. 1, 2012.

[11] R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds., *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, Springer-Verlag, Berlin, Heidelberg, 2008.

[12] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The pascal chime speech separation and recognition challenge," *Computer Speech & Language*, 2012.

[13] A.S. Bregman, *Auditory scene analysis*, MITpress, Cambridge, USA, 1990.

[14] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *Third International Conference on Spoken Language Processing*, 1994.

[15] J. Barker, "Missing data techniques: Recognition with incomplete spectrograms," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and Bhiksha Raj, Eds. Wiley, 2012.

[16] R. Bhiksha and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[17] J. Eggink and G. J. Brown, "A missing feature approach to instrument identification in polyphonic music," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, 2003, pp. 553–556.

[18] D. Giannoulis, A. Klapuri, and M. D. Plumbley, "Recognition of harmonic sounds in polyphonic audio using a missing feature approach," in *submitted to the 38th International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[19] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. of the International Society for Music Information Retrieval (ISMIR)*, 2006, vol. 6, pp. 216–221.

[20] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Tech. Rep., IRCAM, Paris, France, Apr. 2004.

[21] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, no. 1, pp. 5–25, 2005.

[22] T. Virtanen, R. Singh, and Bhiksha Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, 2012.

[23] M.L. Seltzer, B. Raj, and R.M. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[24] A. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 3381–3384.

[25] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *International Conference on Music Information Retrieval*, Baltimore, USA, 2003, pp. 229–230.

[26] F. Opolko and J. Wapnick, *MUMS: McGill University Master Samples*, Montreal, Canada, 1987.

[27] P. Herrera-Boyer, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds., pp. 163–200. Springer, New York, 2006.

[28] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2001.

[29] "freesound.org," Sample numbers (names): 31381 (stall shower), 31487 (bar crowd), 32908 (doing dishes) and 58858 (raindandrumble), (Last accessed: 10/11/2012).