CULTIVATING VOCAL ACTIVITY DETECTION FOR MUSIC AUDIO SIGNALS IN A CIRCULATION-TYPE CROWDSOURCING ECOSYSTEM

Kazuyoshi Yoshii *Hiromasa Fujihara* *Tomoyasu Nakano* *Masataka Goto*

National Institute of Advanced Industrial Science and Technology (AIST)
{k.yoshii, t.nakano, m.goto}@aist.go.jp

ABSTRACT

This paper presents a crowdsourcing-based self-improvement framework of vocal activity detection (VAD) for music audio signals. A standard approach to VAD is to train a vocal-and-non-vocal classifier by using labeled audio signals (training set) and then use that classifier to label unseen signals. Using this technique, we have developed an online music-listening service called Songle that can help users better understand music by visualizing automatically estimated vocal regions and pitches of arbitrary songs existing on the Web. The accuracy of VAD is limited, however, because in general the acoustic characteristics of the training set are different from those of *real* songs on the Web. To overcome this limitation, we adapt a classifier by leveraging vocal regions and pitches corrected by volunteer users. UnlikeWikipedia-type crowdsourcing, our Songle-based framework can *amplify* user contributions: error corrections made for a limited number of songs improve VAD for all songs. This gives better music listening experiences to all users as non-monetary rewards.

Index Terms— Music signal analysis, vocal activity detection, melody extraction, probabilistic models, crowdsourcing

1. INTRODUCTION

Vocal activity detection (VAD) for music audio signals is the basis of a wide range of applications. In retrieval systems, the presence or absence of vocal activity (singing) is one of the most important factors determining a user's preferences. Some people like standard popular songs with vocals and others prefer instrumental pieces without vocals. Music professionals such as disk jockeys and sound engineers often use vocal activity information to efficiently navigate to positions of interest within a target song (*e.g.*, the beginning of singing or of a bridge section played by musical instruments). Accurate VAD is also expected to improve automatic lyric-to-audio synchronization [1, 2] and lyric recognition for music audio signals [3–6].

The major problem of conventional studies on music signal analysis is that almost all methods have been closed in the research community. Although some researchers release source codes for "reproducible research," people who are not researchers cannot enjoy the benefits of the state-of-the-art methods. In addition, we cannot evaluate how well the methods work in the *real* environment. In Japan, for example, numerous original songs composed using the singing-synthesis software called Vocaloid have gained a lot of popularity. Since the acoustic characteristics of synthesized vocals might differ from those of natural human vocals, for those real songs the accuracy of VAD is thought to be limited if the methods are tuned using common music datasets [7, 8] at a laboratory level.

To solve this problem, we have developed a public-oriented online music-listening service called Songle [9] that can assist users to better understand music thanks to the power of music signals analysis. In the current implementation, four kinds of musical elements of arbitrary songs existing on the Web can be estimated: beats, chords, melodies, and structures. Users can enjoy intuitive visualization and sonification of those estimated elements in synchronization with music playback. To estimate main melodies (vocal regions and F0s) of music audio signals, Songle uses VAD and predominant fundamental frequency (F0) estimation methods [10, 11] that can work well for commercial audio recordings of popular music.

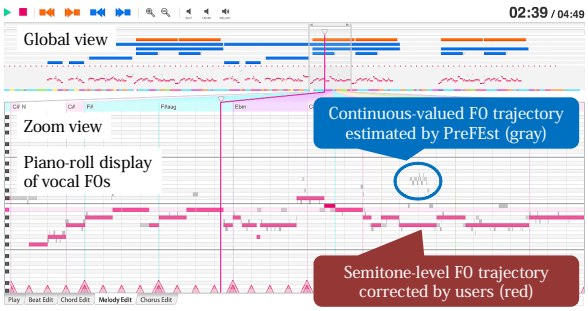

Fig. 1. The melody correction interface of the online music-listening service Songle: Users can correct wrongly-estimated vocal regions and F0s on a Web browser as if they use a MIDI sequencer.

A key feature of Songle[1] is that users can intuitively correct estimation errors on a Web browser. Such voluntary error correction is motivated by prompt feedback of better music-listening experience based on correctly visualized and sonificated musical elements. For example, the melody correction interface is shown in Fig. 1. Note that *true* F0s take continuous values [Hz] and often fluctuate over a semitone because of vibrato, but it is too hard for users to correct estimated F0s precisely. Users are assumed to correct vocal regions at a sixteenth-note level and F0s at a semitone level on an easy-to-use MIDI-sequencer-like interface based on quantized grids.

In this paper we propose a novel crowdsourcing framework that can *cultivate* music signal analysis methods in the real environment by leveraging error corrections made by users. A basic idea for improvingVAD is to use vocal regions and semitone-level F0s specified by users as additional training data. However, the VAD method [10] used in Songle needs *precise* F0s for extracting reliable acoustic features of the main melody. To solve this problem, we re-estimate the F0 at each frame accurately by using a predominant-F0 estimation method [11] that can consider the semitone-level F0 as prior knowl-

This study was supported in part by the JST OngaCREST project.

[1] Songle has officially been open to the public (http::://songle.jp).
A Japanese Vocaloid song composed by talented amateurs:
http://songle.jp/songs/www.youtube.com%2Fwatch%3Fv=PqJNc9KVIZE
A English popular song composed by professional musicians:
http://songle.jp/songs/www.youtube.com%2Fwatch%3Fv=1kz6hNDlEEg

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

edge. Unlike other crowdsourcing services, our framework can *amplify* user contributions. That is, error corrections made for several songs improve VAD for all songs, resulting in positive feedback (better music-listening experiences) to all users. Such non-monetary rewards would motivate users to voluntarily make more corrections in this circulation-type crowdsourcing ecosystem.

2. RELATED WORK

This section introduces several studies on vocal activity detection (VAD) and crowdsourcing for music information processing.

2.1. Vocal Activity Detection and F0 Estimation

Vocal activity detection (VAD) is a typical supervised classification task that aims to detect vocal regions (frames) in music audio signals. A basic approach is to train a binary vocal-and-non-vocal classifier by using frame-level acoustic features extracted from labeled audio signals. This approach was inspired by *voice* activity detection in speech signals for speech recognition [12]. Berenzweig and Ellis [13], for example, extracted phonetic features from music audio signals by using a hidden Markov model (HMM) that was trained using speech signals. Nwe *et al.* [14] tried to attenuate accompanying harmonic sounds by using key information before feature extraction. Lukashevich *et al.* [15] used Gaussian mixture models (GMMs) as a classifier and smoothed the frame-level estimates of class labels by using an autoregressive moving-average (ARMA) filter. Ramona *et al.* [16] used a support vector machine (SVM) as a binary classifier and then used a HMM as a smoother.

Fundamental frequency (F0) of main melodies can be effectively used for improving VAD. Fujihara *et al.* [10, 17], for example, separated main melodies sung by vocalists or played by musical instruments (*e.g.*, solo guitar) from music audio signals by automatically estimating predominant F0s. Although automatic F0 estimation [11] was imperfect, VAD for separated main-melody signals was more accurate than VAD for original music signals. Rao *et al.* [18] took a similar approach based on another F0 estimation method [19]. Both methods used standard GMM-based classifiers.

2.2. Crowdsourcing and Social Annotation

Crowdsourcing is a very powerful tool for gathering a large amount of ground-truth data (for a review see [20]). Recently, Amazon Mechanical Turk (MT) has often been used for music information research. For example, Lee [21] collected subjective judgments about music similarity from MT and needed only 12 hours to collect judgments that took two weeks to collect from experts. Mandel *et al.* [22] showed how social tags for musical pieces crowdsourced from MT could be used for training an autotagger.

There is another kind of crowdsourcing called social annotation. A key feature that would motivate users to make annotations is that annotations made by a user are widely shared among all users (*e.g.*, Wikipedia). Users often want to let others know their favorite items even though they are not monetarily rewarded. In conventional social annotation services, however, improvements based on user contributions are limited to items directly edited by users. To overcome this limitation, an online speech-retrieval service named PodCastle [23] has been developed. In this service, speech signals are automatically transcribed for making text retrieval feasible. A key feature of PodCastle is that users' corrections of transcribed texts are leveraged for improving speech recognition. This leads to better speech retrieval for all users. An online music-listening service named Songle [9] can be regarded as a music version of PodCastle.

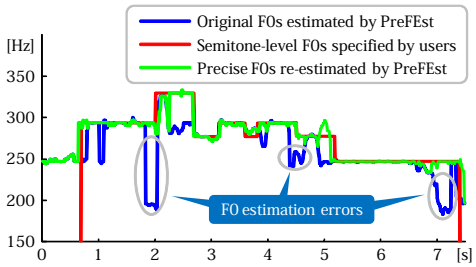

Fig. 2. Comparison of predominant F0 trajectories: Precise F0s can be estimated by using PreFEst [11], which takes into account as prior knowledge the semitone-level F0s specified by users.

3. VOCAL ACTIVITY DETECTION

This section describes a proposed framework that can improve the accuracy of vocal activity detection (VAD) by leveraging the power of crowdsourcing. Our goal is to find vocal regions frommusic audio signals (*i.e.*, to classify frames into vocal and non-vocal classes). In this study we use a competitive VAD method [10] used for singing melody visualization in an online music-listening service Songle [9]. A key feature of this method is to use predominant F0s for extracting acoustic features that represent the timbral characteristics of main melodies. The basic procedure is as follows:

Training phase A classifier based on vocal and non-vocal GMMs is trained using music audio signals with ground-truth annotations (vocal frames and precise vocal F0s in those frames). Because non-vocal frames have no F0 annotations, a predominant F0 estimation method called PreFEst [11] is used for estimating non-vocal F0s in those frames. Spectral-envelope features of main melodies are extracted from vocal and non-vocal frames and then used for training the GMMs.

Classification phase Predominant F0s over all frames are estimated from a target audio signal by using PreFEst. Spectral-envelope features of main melodies are extracted from all frames and then classified by using the trained classifier.

A basic approach to improving VAD is to increase the amount of training data by using online music audio signals annotated by users. Such data can be obtained from Songle, which enables users to correct wrongly-estimated vocal regions and F0s. However, since users are for practical reasons assumed to correct vocal F0s at a semitone level, we cannot extract reliable acoustic features based on *precise* F0s that usually fluctuate over time. To solve this problem, we propose to re-estimate precise F0s by using PreFEst, which as shown in Fig. 2 considers semitone-level F0s as prior knowledge. This is a kind of user-guided F0 estimation.

3.1. Predominant F0 Estimation with Prior Knowledge

To estimate the predominant F0 at each frame, we use a method called PreFEst [11]. The state-of-the-art methods [24, 25] could be used for initial F0 estimation without prior knowledge. To represent the shape of the amplitude spectrum of each frame, PreFEst formulates a probabilistic model consisting of a limited number of parameters. F0 estimation is equivalent to to finding model parameters that maximize the likelihood of the given amplitude spectrum.

3.1.1. Probabilistic Model Formulation

PreFEst tries to learn a probabilistic model that gives the best explanation for the observed amplitude spectrum of each frame. Note that

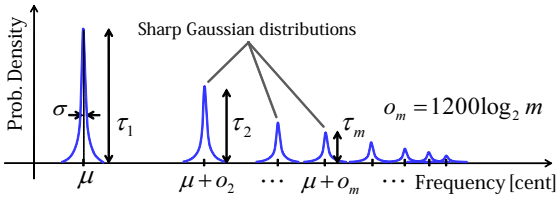

Fig. 3. A constrained GMM representing a harmonic structure

amplitude spectra and harmonic structures are dealt with in the log-frequency domain because the relative positions of harmonic partials are shift-invariant regardless of the F0. Let M be the number of harmonic partials. As shown in Fig. 3, a constrained GMM is used for representing a single harmonic structure as follows:

$$p(x|\mu, \boldsymbol{\tau}) = \sum_{m=1}^{M} \tau_m \mathcal{N}\left(x \middle| \mu + 1200\log_2 m, \sigma^2\right), \quad (1)$$

where x indicates a log-frequency[2], mean μ is the F0 of the harmonic structure, variance σ^2 is the degree of energy diffusion around the F0, and mixing ratio τ_m indicates a relative strength of the m-th harmonic partial ($1 \leq m \leq M$). This means that M Gaussians are located to have harmonic relationships on the log-frequency scale.

As shown in Fig. 4, the amplitude spectrum that might contain multiple harmonic structures is modeled by superimposing all possible harmonic GMMs with different F0s as follows:

$$p(x|\boldsymbol{\tau}, p(\mu)) = \int p(\mu) p(x|\mu, \boldsymbol{\tau}) d\mu, \quad (2)$$

where $p(\mu)$ is a probability distribution of the F0. In this model, $\boldsymbol{\tau}$ and $p(\mu)$ are unknown parameters to be learned (σ^2 is fixed).

3.1.2. Maximum-a-Posteriori Estimation

If prior knowledge is available, it can be taken into account for appropriately estimating $\boldsymbol{\tau}$ and $p(\mu)$ from the given amplitude spectrum [26]. More specifically, prior distributions are given by

$$p(\boldsymbol{\tau}) \propto \exp\left(-\beta_{\boldsymbol{\tau}} \mathcal{D}_{\mathrm{KL}}(\boldsymbol{\tau}_0|\boldsymbol{\tau})\right), \quad (3)$$

$$p(p(\mu)) \propto \exp\left(-\beta_{\mu} \mathcal{D}_{\mathrm{KL}}(p_0(\mu)|p(\mu))\right), \quad (4)$$

where $\mathcal{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence, $\boldsymbol{\tau}_0$ is prior knowledge about the relative strengths of harmonic partials, and $p_0(\mu)$ is prior knowledge about the distribution of the predominant F0. $\beta_{\boldsymbol{\tau}}$ and β_{μ} control how much emphasis is put on those priors.

Those prior distributions have an effect that makes $\boldsymbol{\tau}$ and $p(\mu)$ close to $\boldsymbol{\tau}_0$ and $p_0(\mu)$. Eq. (3) is always considered by setting $\boldsymbol{\tau}_0$ to average relative strengths of harmonic partials. Eq. (4), on the other hand, is taken into account only at vocal frames where semitone-level F0s are given by users. In [26], $p_0(\mu)$ is given by

$$p_0(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2), \quad (5)$$

where μ_0 is a semitone-level F0 and σ_0 is the standard deviation of a precise F0 μ around μ_0 (we set $\sigma_0 = 100$ [cents]).

We then perform maximum-a-posteriori (MAP) estimation of $\boldsymbol{\tau}$ and $p(\mu)$. An objective function to be maximized is given by

$$\int A(x)\left(\log p(x|\boldsymbol{\tau}, p(\mu)) + \log p(\boldsymbol{\tau}) + \log p(p(\mu))\right) dx, \quad (6)$$

[2]Linear frequency f_h in hertz can be converted to log-frequency f_c in cents as follows: $f_c = 1200\log_2(f_h/(440 \times 2^{-4.75}))$.

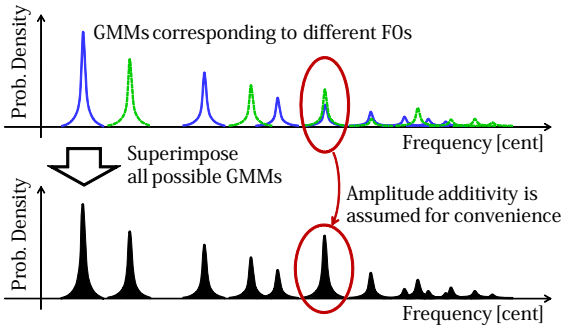

Fig. 4. A probabilistic model for a given amplitude spectrum

where $A(x)$ is the observed amplitude spectrum of the target frame. Since direct maximization of Eq. (6) is analytically intractable, the expectation-maximization (EM) algorithm is used for iteratively optimizing $\boldsymbol{\tau}$ and $p(\mu)$. The predominant F0 is obtained by picking the highest peak from $p(\mu)$. For details see [11] and [26].

3.2. Feature Extraction

To avoid the distortion of acoustic features caused by accompanying instruments, the main melody (not limited to vocal regions) is separated from a target music audio signal. More specifically, we extract a set of harmonic partials at each frame by using an estimated vocal or non-vocal F0 and resynthesize the audio signal by using a well-known sinusoidal synthesis method.

LPC-derived mel-cepstrum coefficients (LPMCCs) are then extracted from the synthesized main melody as acoustic features useful for VAD [17]. The timbral characteristics of speech and singing signals are known to be represented by their spectral envelopes. LPMCCs are mel-cepstrum coefficients of a linear predictive coding (LPC) spectrum. Since cepstrum analysis plays a role of orthogonalization, LPMCCs are superior to the linear predictive coefficients (LPCs) for the classification task. The order of LPMCCs was set to 13.

3.3. Classification

A hidden Markov model (HMM) is used for classifying a feature vector (a set of LPMCCs) of each frame into vocal and non-vocal classes. This HMM consists of vocal and non-vocal GMMs trained using annotated data (musical pieces included in a research-purpose database and online musical pieces annotated by users). To obtain estimates of class labels smoothed over time, the self-transition probabilities ($1.0 - 10^{-40}$) are set to be much larger than the transition probabilities (10^{-40}) between vocal and non-vocal classes. The balance between the hit and correct-rejection rates can be controlled.

3.3.1. Viterbi Decoding

The HMM transitions back and forth between a vocal state s_V and a non-vocal state s_N. Given the feature vectors of a target audio signal $\hat{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots\}$, our goal is to find the most likely sequence of vocal and non-vocal states $\hat{S} = \{s_1, \cdots, s_t, \cdots\}$, *i.e.*,

$$\hat{S} = \underset{S}{\operatorname{argmax}} \sum_t \left(\log p(\boldsymbol{x}_t|s_t) + \log p(s_{t+1}|s_t)\right), \quad (7)$$

where $p(\boldsymbol{x}_t|s_t)$ represents an output probability (vocal or non-vocal GMM) of state s_t, and $p(s_{t+1}|s_t)$ represents the transition probability from state s_t to state s_{t+1}. This decoding problem can be solved efficiently by using the Viterbi algorithm.

The output log-probabilities are given by

$$\log p(\boldsymbol{x}_t|s_V) = \log \mathcal{M}(\boldsymbol{x}_t|\theta_V) - \frac{1}{2}\eta, \quad (8)$$

$$\log p(\boldsymbol{x}_t|s_N) = \log \mathcal{M}(\boldsymbol{x}_t|\theta_N) + \frac{1}{2}\eta, \quad (9)$$

where $\mathcal{M}(\boldsymbol{x}|\theta)$ denotes the likelihood of $\boldsymbol{x}$ in a GMM with parameter θ and η represents a threshold that controls the trad-off between the hit and correct-rejection rates. The parameters of the vocal and non-vocal GMMs, θ_V and θ_N, are trained from LPMCC feature vectors extracted from vocal and non-vocal regions of training data, respectively. We set the number of GMM mixtures to 64.

3.3.2. Threshold Adjustment

The balance between the hit and correct-rejection rates is controlled by changing η in Eqs. (8) and (9). Since the GMM likelihoods are differently distributed for each song, it is hard to decide the universal value of η. Therefore the value of η is adapted to a target audio signal by using a well-known binary discriminant analysis method [27].

4. EVALUATION

This section reports experiments that were conducted for evaluating the improvement of VAD based on crowdsourcing.

4.1. Experimental Conditions

We used two kind of music data. One is a set of 100 songs contained in the RWC Music Database: Popular Music [7] (called RWC data), and the other is a set of 100 "real" musical pieces available on Songle (called Songle data). The RWC data had ground-truth annotations made by experts [8] including precise vocal F0s and regions. On the other hand, the Songle data has partially been annotated by users. Note that users are assumed to correct fluctuating F0s at a semitone level and do nothing for *correctly-estimated* non-vocal regions. In the current Songle interface, we cannot judge whether non-vocal regions that were not corrected by users were actually confirmed to be correct or just unchecked. Therefore in the Songle data the number of non-vocal frames available for training was much smaller than that of annotated vocal frames. We used user annotations as ground truth. A remarkable fact is that there are very few malicious users because Songle is a *non-monetary* crowdsourcing service.

We tested the VAD method [10] in three different ways. To train a classifier, we used only the RWC data (case A) or both the RWC data and the Songle data (cases B and C). In case B, semitone-level F0s given by users were directly used for feature extraction. In case C, on the other hand, precise F0s were re-estimated by using PreFEst that considered semitone-level F0s as prior knowledge.

We measured the accuracy of classification (a rate of correctly-classified frames) on the Songle data. In cases B and C, we conducted 10-fold cross validation, *i.e.*, the RWC data and 90% of the Songle data were used for training and the rest Songle data were used for evaluation. We then performed VAD for 50 songs whose vocals were synthesized by the Vocaloid software. Those songs were chosen from the top ranks in the play-count ranking (not an strictly open evaluation) and were completely annotated by experts.

4.2. Experimental Results

The accuracies of VAD on the Songle data were 66.6%, 67.6%, and 69.6% in cases A, B, and C, respectively. This showed that the proposed crowdsourcing framework (case C) was useful for analyzing real musical pieces outside the laboratory environment. The difference between cases B and C indicated that it was effective to estimate precise F0s before feature extraction by using PreFEst considering semitone-level F0s as prior knowledge. As shown in Table 1 and Table 2, the obtained confusion matrices showed that the number of true negatives (correctly classified non-vocal frames) was increased while the number of true positives (correctly classified vocal frames) was not significantly increased. Note that the vocal GMMs in cases B and C were trained by using plenty of vocal frames in the Songle data. Interestingly, this was useful for preventing non-vocal frames from being misclassified as the vocal class.

Table 1. A confusion matrix obtained in case A (baseline)

Annotation \ Prediction	V	NV
Vocal (V)	**12,347** s	4,086 s
Non-vocal (NV)	2,252 s	**268 s**

Table 2. Confusion matrices obtained in case B and case C

Without precise F0 estimation

	V	NV
V	**12,452 s**	3,981 s
NV	2,152 s	**368 s**

With precise F0 estimation

	V	NV
V	**12,505 s**	3,928 s
NV	1,827 s	**693 s**

There are several reasons that the VAD accuracies on the Songle data were below 70% in this experiment. Firstly, non-vocal frames available for evaluation were much fewer than vocal frames available for evaluation. Secondly, the available non-vocal frames were essentially difficult to be classified because Songle originally misclassified those frames as the vocal class. Thanks to error corrections made by users, such confusing frames could be used for evaluation. Note that the accuracy was 79.6% when we conducted 10-fold class validation on the RWC data. The accuracy on the Vocaloid data, however, was 74.4% when we used only the RWC data for training.

We confirmed that the accuracy on the Vocaloid data was improved to 75.7% by using the Songle data including many Vocaloid songs as additional training data. There is much room for improving VAD. As suggested in [14, 28], it is effective to use a wide range of acoustic features not limited to LPMCCs. It is also important to incrementally cultivate the VAD method by collecting more annotated data from the user-beneficial crowdsourcing framework.

5. CONCLUSION

This paper presented a crowdsourcing-based self-improvement framework of vocal activity detection (VAD) for music audio signals. Our framework trains a better classifier by collecting user-made corrections of vocal F0s and regions from Songle. Since vocal F0s are corrected at a semitone level, we proposed to estimate precise F0s by using as prior knowledge those semitone-level F0s. This enables us to extract reliable acoustic features. The experimental results showed that the accuracy of VAD can be improved by regarding user corrections as additional ground-truth data.

This pioneering work opens up a new research direction. Various kinds of music signal analysis, such as chord recognition and auto-tagging, could be improved by using the power of crowdsourcing. We believe that it is important to design a non-monetary ecosystem, *i.e.*, reward users with the benefits of improved music signal analysis. This could a good incentive to provide high-quality annotations. Songle is a well-designed research platform in which technical improvements are inextricably linked to user contributions.

6. REFERENCES

[1] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy, "Lyrically: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 338–349, 2008.

[2] C. H. Wong, W. M. Szeto, and K. H. Wong, "Automatic lyrics alignment for cantonese popular music," *Multimedia Systems*, vol. 12, no. 4–5, pp. 307–323, 2007.

[3] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010, Article ID 546047.

[4] A. Sasou, M. Goto, S. Hayamizu, and K. Tanaka, "An auto-regressive, non-stationary excited signal parameter estimation method and an evaluation of a singing-voice recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 237–240.

[5] T. Hosoya, M. Suzuki, A. Ito, and S. Makino, "Lyrics recognition from a singing voice based on finite state automaton form music information retrieval," in *International Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 532–535.

[6] C.-K. Wang, R.-Y. Lyu, and Y.-C. Chiang, "An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker," in *European Conference on Speech Communication and Technology (Eurospeech)*, 2003, pp. 1197–1200.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in *International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

[8] M. Goto, "AIST annotation for the RWC music database," in *International Conference on Music Information Retrieval (ISMIR)*, 2006, pp. 359–360.

[9] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano, "Songle: A web service for active music listening improved by user contributions," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2011, pp. 311–316.

[10] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1252–1261, 2011.

[11] M. Goto, "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[12] M. Grimm and K. Kroschel, Eds., *Robust Speech Recognition and Understanding*, I-Tech Education and Publishing, 2007.

[13] A. Berenzweig and D. Ellis, "Locating singing voice segments within music signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001, pp. 119–122.

[14] T. L. Nwe, A. Shenoy, and Y. Wang, "Singing voice detection in popular music," in *ACM Multimedia*, 2004, pp. 324–327.

[15] H. Lukashevich, M. Gruhne, and C. Dittmar, "Effective singing voice detection in popular music using ARMA filtering," in *International Conference on Digital Audio Effects (DAFx)*, 2007, pp. 1–8.

[16] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 1885–1888.

[17] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 638–648, 2010.

[18] V. Rao, C. Gupta, and P. Rao, "Context-aware features for singing voice detection in polyphonic music," in *International Conference on Adaptive Multimedia Retrieval (AMR)*, 2011, pp. 43–57.

[19] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 8, pp. 2145–1254, 2010.

[20] N. Ramzan, F. Dufaux, M. Larson, and K. Clüver, "The participation payoff: Challenges and opportunities for multimedia access in networked communities," in *International Conference on Multimedia Information Retrieval (MIR)*, 2010, pp. 487–496.

[21] J. H. Lee, "Crowdsourcing music similarity judgments using mechanical turk," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2010, pp. 183–188.

[22] M. Mandel, D. Eck, and Y. Bengio, "Learning tags that vary within a song," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2010, pp. 399–404.

[23] M. Goto, J. Ogata, and K. Eto, "A Web 2.0 approach to speech recognition research," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2007, pp. 2397–2400.

[24] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics on Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[25] B. Fuentes, A. Liutkus, R. Badeau, and G. Richard, "Probabilistic model for main melody extraction using constant-Q transform," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 5357–5360.

[26] M. Goto, "A predominant-F0 estimation method for real-world musical audio signals: MAP estimation for incorporating prior knowledge about F0s and tone models," in *Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis (CRAC)*, 2001.

[27] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[28] M. Mauch, H. Fujihara, K. Yoshii, and M. Goto, "Timbre and melody features for the recognition of vocal activity and instrumental solos in polyphonic music," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2011, pp. 233–238.