Real-Time Audio-to-Score Alignment of Singing Voice Based on Melody and Lyric Information

Rong Gong[1,2], *Philippe Cuvillier*[1,2], *Nicolas Obin*[1], *Arshia Cont*[1,2]

[1]IRCAM - UMR STMS IRCAM-CNRS-UPMC
[2]INRIA, MuTant Team-Project
Paris, France

Abstract

Singing voice is specific in music: a vocal performance conveys both music (melody/pitch) and lyrics (text/phoneme) content. This paper aims at exploiting the advantages of melody and lyric information for real-time audio-to-score alignment of singing voice. First, lyrics are added as a separate observation stream into a template-based hidden semi-Markov model (HSMM), whose observation model is based on the construction of vowel templates. Second, early and late fusion of melody and lyric information are processed during real-time audio-to-score alignment. An experiment conducted with two professional singers (male/female) shows that the performance of a lyrics-based system is comparable to that of melody-based score following systems. Furthermore, late fusion of melody and lyric information substantially improves the alignment performance. Finally, maximum a posteriori adaptation (MAP) of the vowel templates from one singer to the other suggests that lyric information can be efficiently used for any singer.

Index Terms: singing voice, real-time audio-to-score alignment, lyrics, spectral envelope, information fusion, singer adaptation.

1. Introduction

Score following is the real-time alignment of incoming audio signals to a symbolic representation of the performance that is available in advance [1]. Score following systems have been at work for automatic musical accompaniment applications since years [2]. The objective is to provide the time position of a performance during real-time execution. Recent score following systems are based on generative probabilistic inference with the implicit assumption that the audio signal is generated by a state-space model representing the symbolic score [3, 4, 1, 5].

A singing voice score contains music and lyric information. Despite the importance of singing voice (especially in popular music repertoire), real-time alignment systems that consider the specificities of the singing voice remain sparse [6, 7]. In particular, real-time alignment systems remain limited to the pitch information derived from the musical score and ignore the lyrics information specific to the singing voice. Alternatively, off-line alignment systems have been developed for audio-to-lyrics alignment inspired by HMM-based speech recognition systems [8, 9, 10]. Also, music and lyric information have been exploited for music information retrieval based on singing voice [11]. These observations encourage the use of lyrics as an alternative source of information to improve the performance of real-time alignment systems for singing voice.

The main objective of this work is to leverage score following for singing voice, by extending the existing *Antescofo* system [5], a template-based hidden semi-Markov model (HSMM) for real-time singing voice audio-to-score alignment. We submit that robust alignment of singing voice must provide specific observation and inference mechanisms that can exploit music and lyric information. The main contributions of this paper are:

- First, we integrate lyrics to the observation mechanism as an alternative source of information (Section 3). The spectral envelope estimated by the True Envelope method [12] is used to construct a set of vowel templates by supervised machine learning, which are then integrated into the alignment system.
- Second, we propose two information fusion strategies to exploit music and lyric information (Section 4). The early fusion performs the fusion of the pitch and vowel templates – accordingly to the source/filter model of voice. The late fusion performs the fusion of pitch and vowel templates observation probabilities.

An objective evaluation of the score-alignment performance for singing voice is reported in Section 5.

2. Real-Time Audio-to-Score Alignment

2.1. Probabilistic Model

Most score following systems are based on a generative probabilistic model which assumes that the audio signal is generated by a hidden state-space model representing the symbolic music score [3, 4, 1, 5]. In particular, the *Antescofo* system is based on a hidden semi-Markov model (HSMM) as defined in [13]. A discrete-time stochastic process $(S_t)_{t\in\mathbb{N}}$ models the hidden position on the score, assumed to be a left-to-right semi-Markov chain, where each state S_t represents one music event in state space J [5]. The observation $(x_1,\ldots,x_\tau)$ consists of fixed-length frames of the acoustic signal generated by the musician, considered as a realization of a stochastic process $(X_t)_{t\in\mathbb{N}}$ that is generated by (S_t). Consequently, audio-to-score alignment consists in finding the most likely state sequence conditionally to the observation sequence. For real-time audio-to-score alignment, (S_t) is sequentially estimated, the current position $\hat{s}_t$ is estimated at time t from past observations only, using the *Forward* recursion:

$$\hat{s}_t = \underset{j\in J}{\operatorname{argmax}} \quad p(S_t = j \mid X_1 = x_1, \ldots, X_t = x_t) \qquad (1)$$

A HSMM assumes that the observation sequence $(X_t)_{t\in\mathbb{N}}$ is conditionally independent on the state sequence $(S_t)_{t\in\mathbb{N}}$.

This project was partially funded by the French ANR INEDIT Project.

Thus, its observation model consists of the specification of the observation probabilities:

$$p(x_t|S_t = j) \overset{\text{def}}{=} p(X_t = x_t|S_t = j) \qquad \forall j \in J \tag{2}$$

2.2. Observation model

The short-term magnitude spectrum SP_t is used as the acoustic observation x_t, and the state space J is deduced from the music score, available in advance (each note is a state).

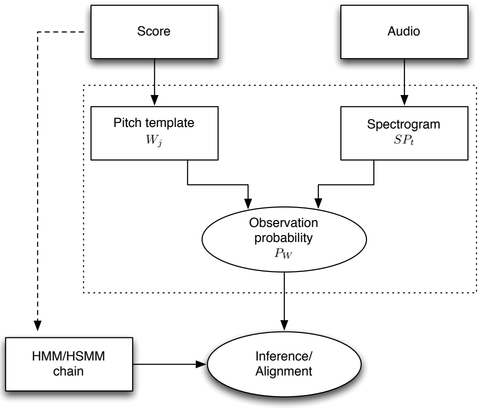

Figure 1: Architecture of the Antescofo system.

2.2.1. *Observation probabilities*

The observation probability is based on a similarity distance of short-term spectrum SP_t with prior pitch spectral templates W_j:

$$p_W(x_t|S_t = j) = \exp[-\beta D(W_j \parallel SP_t)] \tag{3}$$

Here, the similarity measure $D(X \parallel Y)$ is the Kullback-Leibler divergence commonly used for audio-to-music alignment [3, 14]:

$$D(X \parallel Y) \overset{\text{def}}{=} \sum_f X(f) \log \frac{X(f)}{Y(f)}. \tag{4}$$

2.2.2. *Pitch templates*

The pitch template W_j represents the ideal spectral distribution emitted by state j. W_j consists of a mixture of peaks at each harmonics of the fundamental frequency f_0 of state j:

$$W(f) = \sum_{k=1}^{K} e(kf_0)\mathcal{N}(f; kf_0, \sigma^2_{f_0,k}). \tag{5}$$

Each peak is modeled as a Gaussian function whose mean equals the harmonic frequency kf_0 and whose variance σ^2 is constant on the logarithmic scale.

$$\mathcal{N}(f; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6}$$

Spectral envelope $e(kf_0)$ is a decreasing exponential function which approximates spectral density of music instruments.

3. Lyrics Observation Model

The objective of this study is to use lyrics as an alternative observation model, in addition to the melody observation model, for real-time audio-to-score alignment of singing voice.

3.1. Singing Voice and Lyrics

Singing voice is specific in music: a singing voice contains music and lyric information. Thus, music information is necessary but not sufficient for the alignment of singing voice. In particular, lyrics can be used as an alternative source of information (as used for HMM-based audio-to-lyrics alignment [8, 9, 10]) for the real-time audio-to-score alignment of singing voice.

Lyrics conveys a linguistic message, whose smallest unit is the phoneme which is defined by a specific configuration of the vocal tract [15]. In singing voice, the musical message generally prevails over the linguistic message [16]. In particular, vowels carry the melody line (stable part), while consonants constitute perturbations to the melody line (transient part). For instance, vowels represent about 90% of phonation time in opera singing [17]. This motivates the use of vowels for the audio-to-lyrics alignment of singing voice.

3.2. Estimation of Spectral Envelope

3.2.1. *Source/Filter Model*

The source/filter model is a standard representation of a speech signal:

$$X(f) = S(f) \times H(f) \tag{7}$$

where $S(f)$ is the frequency response of the glottal source excitation, $H(f)$ is the frequency response of the vocal-tract filter.

The source excitation $S(f)$ encodes the pitch information (music), and the vocal tract $H(f)$ encodes the phoneme information (lyrics). The spectral envelope is commonly used to estimate of frequency response of the vocal-tract filter.

3.2.2. *True Envelope*

The cepstrum [18] is a wide-spread representation used for source/filter deconvolution, and spectral envelope estimation [19, 20] (among other existing representations, e.g., Linear Predictive Coding LPC [21], and with extension to Mel-Frequency Cepstral Coefficients (MFCC) [22]).

The True Envelope (TE) is an iterative method for cepstrum-based spectral envelope estimation [12, 23] (Figure 2). At iteration i, the log-amplitude spectral envelope $E_i(f)$ of the log-amplitude spectrum $X_i(f)$ is given by

$$E_i(f) = c(0) + 2\sum_{p=1}^{P} c(p)\cos(2\pi fp) \tag{8}$$

where $c(p)$ denotes the $p - th$ cepstral coefficient, and P the number of cepstral coefficients used for the spectral envelope estimation.

The TE method iterates as follows:

1. Initialize “target” spectrum and spectral envelope:

   $X_0(f) = \log(|X(f)|)$
   $V_0(f) = -\infty, \qquad \forall f$

2. Update “target” spectrum amplitude at iteration i:

   $E_i(f) = \max(X_{i-1}(f), E_{i-1}(f)), \qquad \forall f$

3. Update cepstrum of “target” spectrum $X_i(f)$, and corresponding spectral envelope $E_i(f)$.

Steps 2 and 3 are repeated until the following criterion of convergence is reached:

$$X_i(f) - E_i(f) \leq \theta, \forall f \tag{9}$$

A typical value of θ is the one that corresponds to 2 dB.

Additionally, the optimal cepstrum order $\hat{P}$ for the estimation of the spectral envelope (thus, source/filter separation) can be directly derived as [20]:

$$\hat{P} = \frac{F_s}{2F_0}$$

where F_s is the sampling frequency of the signal, and F_0 is the fundamental frequency of the signal.

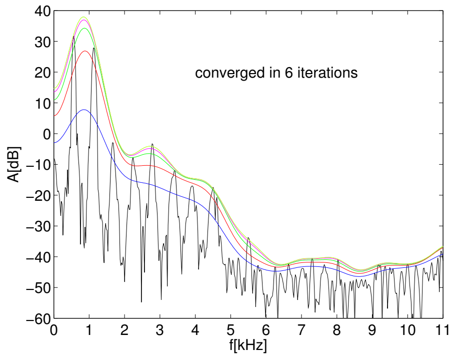

Figure 2: True Envelope estimation of the spectral envelope. In black, original spectrum; in blue, cepstrum estimation of the spectral envelope; other colors, iterative estimations of the spectral envelope.

3.3. Integration of Lyric Information

The template-based observation model described in section 2.2 is frequently used by score following systems. While most templates design are based on heuristic choices like the the harmonic mixture of Gaussians in equation (5) [3, 5, 24, 25], some systems adopt machine learning for templates design [1, 26]. Here, machine learning is adopted to design vowel templates.

3.3.1. Observation model

The observation model used for lyrics alignment replaces pitch-based observations with vowel-based observations, and pitch templates with vowel templates, while assuming the same observation probability function as the one in section 2.2.1. The observation probability is thus defined as the similarity distance of short-term spectral envelope TE_t with prior vowel templates V_j:

$$p_V(x_t|S_t = j) = \exp[-\beta D(V_j \parallel TE_t)]. \tag{10}$$

3.3.2. Vowel Templates

Supervised machine learning is used to estimate vowel templates from a training set of manually labeled recordings. Like [26], we consider the Maximum Likelihood Estimator (MLE). The j^{th} vowel template V_j is determined as explained in [14, Theorem 2]:

$$V_j(f) = \nabla F \left(\frac{1}{N_j} \sum_{n=1}^{N_j} TE_n^{(j)}(f) \right) \tag{11}$$

where $TE_n^{(j)}$ is the n^{th} true envelope frame corresponding to vowel j, N_j is the total number of frames corresponding to vowel j, and ∇F is the gradient of the cumulative function F associated to KL divergence. Here, the set of vowel templates is constructed for each singer separately.

3.3.3. Adaptation of Vowel Templates

The main issue with speech and singing is the acoustic variability between singers. Consequently, the vowel templates of one singer may be significantly different to those of another singer. In order to exploit the lyric information regardless to the singer, one must adapt the vowel templates of a singer to a small amount of observations of another singer. This can be processed by Maximum A Posteriori (MAP) adaptation [27]:

$$\widehat{V_j}(f)^{(k)} = \alpha E(\mathbf{TE}(f)^{(j)}) + (1-\alpha)\widehat{V_j}(f)^{(k-1)} \tag{12}$$

where: k is the iteration of the MAP adaptation, $\widehat{V_j}(f)$ the $j-th$ vowel template, $E(\mathbf{TE}(f)^{(j)})$ the expectation of all true envelope observations of the $i-th$ vowel, and α the adaptation coefficient.

4. Fusion of Melody and Lyric information

To exploit melody and lyric information, two information fusion are investigated: early fusion of observations and late fusion of observation probabilities.

4.1. Early fusion

The early fusion strategy consists of fusing the observations of pitch and vowel templates, as inspired by the source/filter model. The strategy consists of merging pitch W_j and vowel V_j templates into a single template T_j. This template is obtained by pointwise spectral multiplication:

$$T_j^{\text{fusion}}(f) = W_j(f) \times V_j(f) \tag{13}$$

Then, the observation probability is computed by comparing the short-term spectrum S_t with the template $T_j^{\text{fusion}}(f)$, as defined in equation (3).

4.2. Late fusion

The late fusion strategy consists of fusing the observation probabilities of pitch and vowel templates. First, probabilities of pitch p_W and vowel p_V templates are computed as previously defined in equations (3) and (10). Then, the fused observation probabilities p_{fusion} are obtained by using to following additive mixture:

$$p_{\text{fusion}} = \frac{p_W + p_V}{2} \tag{14}$$

Here, the additive operator allows a stronger observation probability to compensate a weak one, which improves the alignment robustness in case where pitch or vowel information is not reliable.

5. Experiment

5.1. Singing Database

The evaluation database used for this experiment contains audio recordings of French popular songs sung by two professional singers (male/female), their music and lyrics scores, and manual alignments. Manual alignments, used as a reference, is the indication of the attack time of each music event in the score – a new event is defined as a change of pitch and/or vowel. We use the X-SAMPA characters to label French vowels [28] (2, 9, 9∼, @, E, O, A, a, a∼, e, e∼, i, o, o∼, u, y.). Then, the music score representation is then extended by appending the X-SAMPA label to each musical event.

For each singer, the database is split into train and test databases.

- train database: the train database contains 16 vowels with 10 instances each, sung with a constant pitch;
- test database: the test database contains around 8 popular songs, around 10 mn. in total, and around 1000 musical events;

All audio were recorded in professional studio with lossless encoding (48 kHz, 16 bits).

5.2. Experimental Setups

The experiment compares four alignment strategies: using **pitch** information only, **vowel** information only, and using **early** and **late** fusion of pitch and vowel information. Additionally, same-singer and cross-singers performances are compared, with and without MAP adaptation of the vowel templates.

- same-singer: vowel templates are constructed from the training database of a singer, and then used for alignment on the test database of the same singer;
- cross-singers: vowel templates are constructed from the training database of a singer, and then used for alignment on the test database of the other singer. MAP adaptation is optionally used to adapt the vowel templates of a singer with respect to the training database of the other singer;

The evaluation metrics follow the international MIREX campagne for real-time music alignment as described in [29] and using three basic *event metrics*:

Error $e_i = |t_i^e - t_i^r|$ is defined as the absolute time lapse between the alignment positions of corresponding events in the annotation t_i^r and the estimated alignment time t_i^e for score events i.

Misaligned notes are events in the score that are recognized but whose absolute error e_i to the reference alignment is greater than θ_e (here, $\theta_e = 300$ ms).

Missed notes are events that are not recognized.

The *assessment metrics* used to measure the quality of the alignment are then: the **average error**, the **misalign rate**, and the **miss rate** which are simply deduced from the corresponding event metrics (see [29] for further details). In this paper, the assessment metrics are computed for each audio recording, and then averaged to provide performance statistics over all audio recordings.

5.3. Results

Table 1 presents the performance obtained by the four strategies for same-singer alignment. First, the vowel information only has comparable alignment performance compared to the pitch information only (slightly lower for misalign rate and miss rate). This confirms the importance of the vowel information for singing voice alignment. Then, the late fusion strategy significantly improves the alignment performance compared to the standard alignment strategy (by 3.89% for misalign rate and by 1.78% for miss rate, compared to the pitch information). The early fusion strategy does not however improve the alignment performance: the fused template accumulates the individual errors of pitch and vowel templates. Beside, this shows that the adequate fusion of pitch and vowel information can substantially improve the performance of score following systems for singing voice.

STRATEGY	Avg. error (ms)	Misal. rate %	Miss rate %
PITCH	75.8 (2.8)	7.9 (4.2)	2.7 (1.5)
VOWEL	84.0 (2.8)	7.4 (2.5)	3.6 (1.7)
EARLY	68.8 (2.7)	7.9 (3.9)	4.1 (1.8)
LATE	**67.8** (2.4)	**4.0** (2.5)	**0.9** (0.4)

Table 1: Mean performance (and 95% confidence interval) for the four strategies for same-singer alignment.

Table 2 presents the performance obtained by the four strategies for cross-singers alignment. First, the use of vowel templates of a singer for cross-singer alignment seriously degrades the alignment performance of the system. This was expected: the vowel templates of a singer cannot be used as a singer-independent model for singing voice alignment, since vowel templates may vary significantly from one singer to the other. Second, the MAP adaptation of the vowel templates from one singer to the other tends to similar alignment performance than the singer-dependent vowel templates. This indicates that the the vowel templates of a singer can be efficiently adapted to another singer, so that the lyric information can be exploited for any singer, with a reasonable amount of recordings of the singer.

STRATEGY	Avg. error (ms)	Misal. rate %	Miss rate %
PITCH	75.8 (2.8)	7.9 (4.2)	2.7 (1.5)
VOWEL W/O MAP	92.2 (3.0)	11.3 (5.3)	7.8 (4.8)
VOWEL W MAP	79.4 (2.7)	6.8 (3.4)	3.1 (2.2)
EARLY W/O MAP	73.6 (2.8)	10.1 (3.9)	5.0 (2.3)
EARLY W MAP	69.1 (2.7)	7.9 (3.8)	4.2 (2.0)
LATE W/O MAP	73.4 (2.6)	5.0 (3.9)	2.0 (1.8)
LATE W MAP	**68.1** (2.4)	**4.3** (2.6)	**1.2** (0.8)

Table 2: Mean performance (and 95% confidence interval) for the four strategies for cross-singers alignment.

6. Conclusion

This paper introduced the use of lyric information in addition to melody information for the real-time score-following of singing voice - through the construction of vowel templates. An objective evaluation showed that the performance of lyric information only is comparable to that of state-of-the-art music score following systems. Furthermore, the late fusion of melody and lyrics observation probabilities substantially improved the alignment performance. Finally, the adaptation of vowel templates from one singer to the other showed that lyric information can be exploited efficiently to any singer. This constitutes a preliminary advance towards the combination of Automatic Speech Recognition (ASR) with score following probabilistic models. Further research will investigate advanced fusion strategies of melody and lyric information, and the on-line adaptation of lyrics templates.

7. References

[1] A. Cont, "Realtime Audio to Score Alignment for Polyphonic Music Instruments, using Sparse Non-Negative Constraints and Hierarchical HMMs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, Toulouse, France, May 2006.

[2] R. B. Dannenberg and C. Raphael, "Music Score Alignment and Computer Accompaniment," *Communications of ACM*, vol. 49, no. 8, pp. 38–43, 2006.

[3] C. Raphael, "Automatic Segmentation of Acoustic Musical Signals using Hidden Markov Models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 4, pp. 360–370, 1999.

[4] N. Orio and F. Déchelle, "Score Following Using Spectral Analysis and Hidden Markov Models," in *International Computer Music Conference (ICMC)*, Havana, Cuba, 2001.

[5] A. Cont, "A Coupled Duration-Focused Architecture for Real-Time Music-to-Score Alignment," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 974–987, 2010.

[6] M. Puckette, "Score Following Using the Sung Voice," in *International Computer Music Conference (ICMC)*, 1995, pp. 199–200.

[7] A. Loscos, P. Cano, and J. Bonada, "Low-Delay Singing Voice Alignment to Text," in *Proceedings of the ICMC*, 1999.

[8] A. Mesaros and T. Virtanen, "Automatic Recognition of Lyrics in Singing," *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[9] H. Fujihara, M. Goto, J. Ogata, and H. 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, October 2011.

[10] M. Mauch, H. Fujihara, and M. Goto, "Integrating Additional Chord Information Into HMM-Based Lyrics-to-Audio Alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 200–210, Jan 2012.

[11] T. Wang, D.-J. Kim, K.-S. Hong, and J.-S. Youn, "Music Information Retrieval System Using Lyrics and Melody Information," in *Information Processing, 2009. APCIP 2009. Asia-Pacific Conference on*, vol. 2, July 2009, pp. 601–604.

[12] A. Röbel and X. Rodet, "Efficient Spectral Envelope Estimation and its Application to Pitch Shifting and Envelope Preservation," in *International conference on Digital Audio Effects (DAFx)*, Madrid, Spain, 2005.

[13] Y. Guédon, "Hidden Hybrid Markov/Semi-Markov Chains," *Computational Statistics and Data Analysis*, vol. 49, pp. 663–68, 2005.

[14] A. Cont, S. Dubnov, and G. Assayag, "On the Information Geometry of Audio Streams With Applications to Similarity Computing," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 837–846, 2011.

[15] C. Gussenhoven and H. Jacobs, *Understanding Phonology*, 2nd ed. London: Hodder Arnold, New York: Oxford University press, 2005.

[16] J. Ginsborg, "The Influence of Interactions between Music and Lyrics: What Factors Underlie the Intelligibility of Sung Text?" *Empirical Musicology Review*, vol. 9, no. 1, pp. 21–24, 2014.

[17] N. S. Di Carlo, "Effect of Multifactorial Constraints on Intelligibility of Opera Singing (II)," *Journal of Singing*, no. 63, 2007.

[18] A. Oppenheim, "Speech Analysis-Synthesis System Based on Homorphic Filtering," *The Journal of the Acoustical Society of America*, vol. 45, no. 1, pp. 458–465, 1969.

[19] S. J. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[20] F. Villavicencio, A. Röbel, and X. Rodet, "Applying Improved Spectral Modeling for High Quality Voice Conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[21] J. Makhoul, "Linear Prediction: A Tutorial Review," in *Proceedings of the IEEE*, vol. 63, no. 4, 1975, pp. 561–580.

[22] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 28, no. 4, p. 357–366, 1980.

[23] F. Villavicencio, A. Röbel, and X. Rodet, "Improving LPC Spectral Envelope Extraction Of Voiced Speech By True-Envelope Estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[24] C. Raphael, "Aligning Music Audio with Symbolic Scores using a Hybrid Graphical Model," *Machine Learning*, vol. 65, no. 2-3, pp. 389–409, 2006.

[25] T. Otsuka, K. Nakadai, T. Takahashi, T. Ogata, and H. G. Okuno, "Real-time Audio-to-Score Alignment Using Particle Filter for Coplayer Music Robots," *EURASIP Journal of Advances in Signal Processing*, vol. 2011, pp. 1–13, 2011.

[26] C. Joder, S. Essid, and G. Richard, "Learning Optimal Features for Polyphonic Audio-to-Score Alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

[27] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[28] J. Wells. Computer-Coding the IPA: a Proposed Extension of SAMPA. Accessed: 2014-07-18. [Online]. Available: http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm

[29] A. Cont, D. Schwarz, N. Schnell, and C. Raphael, "Evaluation of Real-Time Audio-to-Score Alignment," in *International Symposium on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.