SPEAKER IDENTIFICATION USING GAUSSIAN MIXTURE MODELS BASED ON MULTI-SPACE PROBABILITY DISTRIBUTION

Chiyomi Miyajima[†], *Yosuke Hattori*[†,*], *Keiichi Tokuda*[†], *Takashi Masuko*[‡], *Takao Kobayashi*[‡], *and Tadashi Kitamura*[†]

[†] Nagoya Institute of Technology, Nagoya 466-8555, Japan
[‡] Tokyo Institute of Technology, Yokohama 226-8502, Japan
[*] Present address: DENSO Corporation, Kariya 448-8661, Japan
[†]{chiyomi, tokuda, kitamura}@ics.nitech.ac.jp, [‡]{masuko, tkobayas}@ip.titech.ac.jp, [*] hattori@rd.denso.co.jp

ABSTRACT

This paper presents a new approach to modeling speech spectra and pitch for text-independent speaker identification using Gaussian mixture models based on multi-space probability distribution (MSD-GMM). The MSD-GMM allows us to model continuous pitch values for voiced frames and discrete symbols representing unvoiced frames in a unified framework. Spectral and pitch features are jointly modeled by a two-stream MSD-GMM. We derive maximum likelihood (ML) estimation formulae for the MSD-GMM parameters, and the MSD-GMM speaker models are evaluated for text-independent speaker identification tasks. Experimental results show that the MSD-GMM can efficiently model spectral and pitch features of each speaker and outperforms conventional speaker models.

1. INTRODUCTION

Gaussian mixture models (GMMs) have been successfully applied to speaker modeling in text-independent speaker identification [1]. Such identification systems mainly use spectral features represented by cepstral coefficients as speaker features. Pitch features as well as spectral features contain much speaker specific information [2, 3]. However, most of speaker recognition studies in recent years have focused on using only spectral features. The main reasons for this are i) the use of pitch features alone could not give enough recognition performance and ii) pitch values are not defined in unvoiced segments and this complicates speaker modeling and feature integration.

Several works have reported that speaker recognition accuracy can be improved by the use of pitch features in addition to spectral features [4, 5, 6, 7]. There are essentially two approaches to integrating spectral and pitch information: i) two separate models are used for spectral and pitch features and their scores are combined [4, 5], ii) two separate models for voiced and unvoiced parts are trained and their scores are combined [6, 7]. In [7], two separate GMMs are used, where the input observations are concatenations of cepstral coefficients and $\log F_0$ for voiced frames and cepstral coefficients alone for unvoiced frames. Since the probability distribution of the conventional GMM is defined on a single vector space, these two kinds of vectors require their respective models.

A part of this work was supported by Research Fellowships from the Japan Society for the Promotion of Science for Young Scientists, No. 199808177.

In this paper a new speaker modeling technique using a GMM based on multi-space probability distribution (MSD) [8] is introduced. The MSD-GMM allows us to model feature vectors with variable dimensionality including zero-dimensional vectors, i.e., discrete symbols. Consequently, continuous pitch values for voiced frames and discrete symbols representing "unvoiced" can be modeled using an MSD-GMM in a unified framework, and spectral and pitch features are jointly modeled by a multi-stream MSD-GMM, i.e., each speaker is modeled by a single statistical model. We derive maximum likelihood (ML) estimation formulae and evaluate the MSD-GMM speaker models for text-independent speaker identification tasks comparing with conventional GMM speaker models.

The rest of the paper is organized as follows. In Section 2, we introduce a speaker modeling technique based on MSD-GMM. Section 3 presents the ML estimation procedure for MSD-GMM parameters. Section 4 reports experimental results, and Section 5 gives conclusions and future works.

2. MULTI-STREAM MSD-GMM

2.1. Likelihood Calculation

Let us assume that a given observation $\boldsymbol{o}_t$ at time t consists of S information sources (streams). The s-th stream $\boldsymbol{o}_{ts}$ has a set of space indices X_{ts} and a random vector with variable dimensionality $\boldsymbol{x}_{ts}$, that is

$$\boldsymbol{o}_t = (\boldsymbol{o}_{t1}, \boldsymbol{o}_{t2}, \dots, \boldsymbol{o}_{tS}), \tag{1}$$

$$\boldsymbol{o}_{ts} = (X_{ts}, \boldsymbol{x}_{ts})\,. \tag{2}$$

Note here that X_{ts} is a subset of all possible space indices $\{1, 2, \dots, G_s\}$, and all the spaces represented by the indices in X_{ts} have the same dimensionality as $\boldsymbol{x}_{ts}$.

We define the output probability distribution of an S-stream MSD-GMM λ for $\boldsymbol{o}_t$ as

$$b(\boldsymbol{o}_t \mid \lambda) = \sum_{m=1}^{M} c_m \prod_{s=1}^{S} p_{ms}(\boldsymbol{o}_{ts}), \tag{3}$$

where c_m is the mixture weight for the m-th mixture component. The observation probability of $\boldsymbol{o}_{ts}$ for mixture m is given by the multi-space probability distribution (MSD) [8]:

$$p_{ms}(\boldsymbol{o}_{ts}) = \sum_{g \in X_{ts}} w_{msg} \mathcal{N}_{msg}^{D_{sg}}(\boldsymbol{x}_{ts}), \tag{4}$$

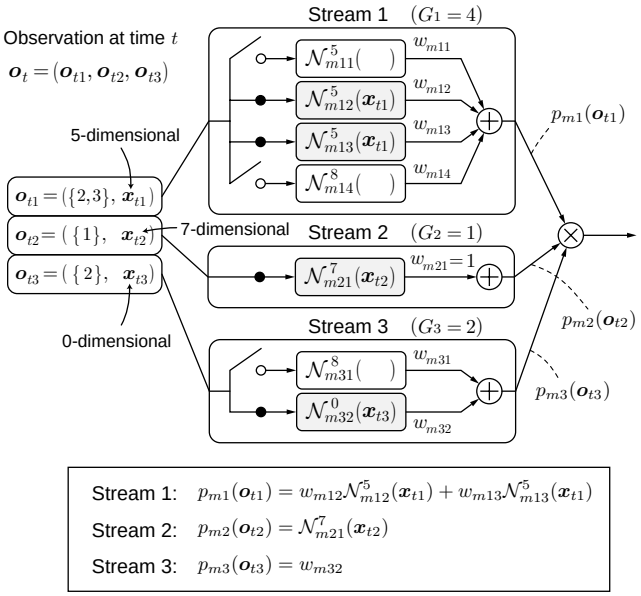

Fig. 1. An example of the m-th mixture component of a three-stream MSD-GMM.

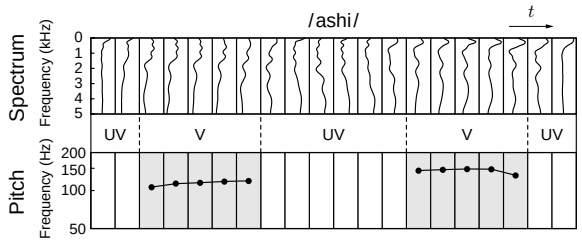

Fig. 2. An example of spectral and pitch sequences of a word "/ashi/" spoken by a male speaker.

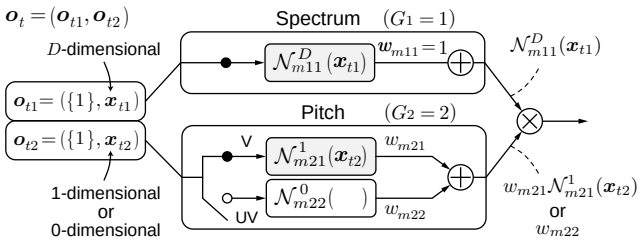

Fig. 3. The m-th mixture component in a two-stream MSD-GMM based on spectra and pitch.

where w_{msg} is the weight for the g-th vector space of the s-th stream and $\mathcal{N}^{D_{sg}}_{msg}(\cdot)$ is the D_{sg}-variate Gaussian function with mean vector $\boldsymbol{\mu}_{msg}$ and covariance matrix $\boldsymbol{\Sigma}_{msg}$ (for the case $D_{sg} > 0$). For simplicity of notation, we define $\mathcal{N}^0_{msg}(\cdot) \equiv 1$ (for the case $D_{sg} = 0$). Note here that the multi-space probability distribution (MSD) is equivalent to continuous probability distribution and discrete probability distribution when $D_{sg} \equiv n > 0$ and $D_{sg} \equiv 0$, respectively. Also, an MSD-GMM is assumed to be a generalized GMM, which includes the traditional GMM as a special case when $S = 1$, $G_1 = 1$, and $D_{11} > 0$.

For an observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$, the likelihood of MSD-GMM λ is given by

$$P(\boldsymbol{O} \mid \lambda) = \prod_{t=1}^{T} b(\boldsymbol{o}_t \mid \lambda). \tag{5}$$

Figure 1 illustrates an example of the m-th mixture component of a three-stream MSD-GMM ($S = 3$). The sample space of the first stream consists of four spaces ($G_1 = 4$), among which the second and third spaces are triggered by the space indices and $p_{m1}(\boldsymbol{o}_{t1})$ becomes the sum of the two weighted Gaussians. The second stream has only one space ($G_2 = 1$) and always outputs its Gaussian as $p_{m2}(\boldsymbol{o}_{t2})$. The third stream consists of two spaces ($G_3 = 2$), where a zero-dimensional space is selected, and outputs its space weight w_{m32} (a discrete probability) as $p_{m3}(\boldsymbol{o}_{t3})$.

2.2. Speaker Modeling Based on MSD-GMM

Figure 2 shows an example of spectral and pitch sequences of a Japanese word "/ashi/" spoken by a Japanese male speaker. Generally, spectral features are represented by multi-dimensional vectors of cepstral coefficients with continuous values. On the other hand, pitch features are represented by one-dimensional continuous values of log fundamental frequencies ($\log F_0$) in voiced frames and discrete symbols representing "unvoiced" in unvoiced frames because pitch values are defined only in voiced segments.

As shown in Fig. 3, each speaker can be modeled by a two-stream MSD-GMM ($S = 2$); the first stream is for the spectral feature and the second stream is for the pitch feature. The spectral stream has a D-dimensional space ($G_1 = 1$), and the pitch stream has two spaces ($G_2 = 2$): a one-dimensional space and a zero-dimensional space for voiced and unvoiced parts, respectively.

3. ML-ESTIMATION FOR MSD-GMM

In a similar way to the ML-estimation procedure in [8], $P(\boldsymbol{O} \mid \lambda)$ is increased by iterating the maximization of an auxiliary function $Q(\lambda', \lambda)$ over λ to improve current parameters λ' based on the expectation maximization (EM) algorithm.

3.1. Definition of Q-Function

The log-likelihood of λ for an observation sequence $\boldsymbol{O}$, a sequence of mixture components $\boldsymbol{i}$ and a sequence of space indices $\boldsymbol{l}$ can be written as

$$\log P(\boldsymbol{O}, \boldsymbol{i}, \boldsymbol{l} \mid \lambda) = \sum_{t=1}^{T} \log c_{i_t} + \sum_{t=1}^{T} \sum_{s=1}^{S} \log w_{i_t s l_{ts}} + \sum_{t=1}^{T} \sum_{s=1}^{S} \log \mathcal{N}^{D_{s l_{ts}}}_{i_t s l_{ts}}(\boldsymbol{x}_{ts}), \tag{6}$$

where

$$\boldsymbol{i} = (i_1, i_2, \ldots, i_T), \tag{7}$$

$$\boldsymbol{l} = (\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T), \tag{8}$$

$$\boldsymbol{l}_t = (l_{t1}, l_{t2}, \ldots, l_{tS}). \tag{9}$$

Hence the Q-function is defined as

$$
\begin{aligned}
Q(\lambda', \lambda) &= \sum_{\text{all } \boldsymbol{i},\boldsymbol{l}} P(\boldsymbol{i},\boldsymbol{l} \mid \boldsymbol{O}, \lambda') \log P(\boldsymbol{O},\boldsymbol{i},\boldsymbol{l} \mid \lambda) \\
&= \sum_{\text{all } \boldsymbol{i},\boldsymbol{l}} P(\boldsymbol{i},\boldsymbol{l} \mid \boldsymbol{O}, \lambda') \sum_{t=1}^{T} \log c_{i_t} \\
&+ \sum_{\text{all } \boldsymbol{i},\boldsymbol{l}} P(\boldsymbol{i},\boldsymbol{l} \mid \boldsymbol{O}, \lambda') \sum_{t=1}^{T} \sum_{s=1}^{S} \log w_{i_t s l_{ts}} \\
&+ \sum_{\text{all } \boldsymbol{i},\boldsymbol{l}} P(\boldsymbol{i},\boldsymbol{l} \mid \boldsymbol{O}, \lambda') \sum_{t=1}^{T} \sum_{s=1}^{S} \log \mathcal{N}^{D_{s l_{ts}}}_{i_t s l_{ts}}(\boldsymbol{x}_{ts}) \\
&= \sum_{m=1}^{M} \sum_{t=1}^{T} P(i_t = m \mid \boldsymbol{O}, \lambda') \log c_m \\
&+ \sum_{m=1}^{M} \sum_{s=1}^{S} \sum_{g=1}^{G_s} \sum_{t \in T(\boldsymbol{O},s,g)} \\
&\qquad P(i_t = m, l_{ts} = g \mid \boldsymbol{O}, \lambda') \log w_{msg} \\
&+ \sum_{m=1}^{M} \sum_{s=1}^{S} \sum_{g=1}^{G_s} \sum_{t \in T(\boldsymbol{O},s,g)} \\
&\qquad P(i_t = m, l_{ts} = g \mid \boldsymbol{O}, \lambda') \log \mathcal{N}^{D_{sg}}_{msg}(\boldsymbol{x}_{ts}), \qquad (10)
\end{aligned}
$$

where

$$
T(\boldsymbol{O}, s, g) = \{t \mid g \in X_{ts}\}. \qquad (11)
$$

3.2. Maximization of Q-Function

The first two terms of (10) have the form $\sum_{i=1}^{N} u_i \log y_i$, which attains a global maximum at the single point

$$
y_i = \frac{u_i}{\sum_{j=1}^{N} u_j}, \quad \text{for} \quad i = 1, 2, \ldots, N, \qquad (12)
$$

under the constraints $\sum_{i=1}^{N} y_i = 1$ and $y_i \geq 0$. The maximization of the first term of (10) leads to the re-estimate of c_m:

$$
\begin{aligned}
c_m &= \frac{\sum_{t=1}^{T} P(i_t = m \mid \boldsymbol{O}, \lambda')}{\sum_{m=1}^{M} \sum_{t=1}^{T} P(i_t = m \mid \boldsymbol{O}, \lambda')} \\
&= \frac{1}{T} \sum_{t=1}^{T} P(i_t = m \mid \boldsymbol{O}, \lambda') \\
&= \frac{1}{T} \sum_{t=1}^{T} \gamma'_t(m), \qquad (13)
\end{aligned}
$$

where $\gamma_t(m)$ is the posterior probability of being in the m-th mixture component at time t, that is

$$
\gamma_t(m) = P(i_t = m \mid \boldsymbol{O}, \lambda) = \frac{c_m \prod_{s=1}^{S} p_{ms}(\boldsymbol{o}_{ts})}{b(\boldsymbol{o}_t)}. \qquad (14)
$$

Similarly, the second term is maximized as

$$
w_{msg} = \frac{\sum_{t \in T(\boldsymbol{O},s,g)} \xi'_{ts}(m, g)}{\sum_{l=1}^{G_s} \sum_{t \in T(\boldsymbol{O},s,l)} \xi'_{ts}(m, l)}, \qquad (15)
$$

where $\xi_{ts}(m, g)$ is the posterior probability of being in the g-th space of stream s in the m-th mixture component at time t:

$$
\begin{aligned}
\xi_{ts}(m, g) &= P(i_t = m, l_{ts} = g \mid \boldsymbol{O}, \lambda) \\
&= P(i_t = m \mid \boldsymbol{O}, \lambda) P(l_{ts} = g \mid i_t = m, \boldsymbol{O}, \lambda) \\
&= \gamma_t(m) \frac{w_{msg} \mathcal{N}^{D_{sg}}_{msg}(\boldsymbol{x}_{ts})}{p_{ms}(\boldsymbol{o}_{ts})}. \qquad (16)
\end{aligned}
$$

The third term is maximized by solving following equations:

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}_{msg}} \sum_{t \in T(\boldsymbol{O},s,g)} P(i_t = m, l_{ts} = g \mid \boldsymbol{O}, \lambda') \\
&\qquad \cdot \log \mathcal{N}^{D_{sg}}_{msg}(\boldsymbol{x}_{ts}) = \boldsymbol{0}, \qquad (17) \\
&\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}_{msg}} \sum_{t \in T(\boldsymbol{O},s,g)} P(i_t = m, l_{ts} = g \mid \boldsymbol{O}, \lambda') \\
&\qquad \cdot \log \mathcal{N}^{D_{sg}}_{msg}(\boldsymbol{x}_{ts}) = \boldsymbol{0}, \qquad (18)
\end{aligned}
$$

resulting in

$$
\boldsymbol{\mu}_{msg} = \frac{\sum_{t \in T(\boldsymbol{O},s,g)} \xi'_{ts}(m, g) \boldsymbol{x}_{ts}}{\sum_{l=1}^{G_s} \sum_{t \in T(\boldsymbol{O},s,l)} \xi'_{ts}(m, l)}, \qquad (19)
$$

$$
\boldsymbol{\Sigma}_{msg} = \frac{\sum_{t \in T(\boldsymbol{O},s,g)} \xi'_{ts}(m, g)(\boldsymbol{x}_{ts} - \boldsymbol{\mu}_{msg})(\boldsymbol{x}_{ts} - \boldsymbol{\mu}_{msg})^{\top}}{\sum_{l=1}^{G_s} \sum_{t \in T(\boldsymbol{O},s,l)} \xi'_{ts}(m, l)}. \qquad (20)
$$

The re-estimation is repeated iteratively using λ in place of λ' and the final result is an ML estimation of the MSD-GMM.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Conditions

First, a speaker identification experiment was carried out using the ATR Japanese speech database. We used word data spoken by 80 speakers (40 males and 40 females). Phonetically-balanced 216 words are used for training each speaker model, and 520 common words per speaker are used for testing. The number of tests was 41600 in total.

Second, to evaluate the robustness of the MSD-GMM speaker model against inter-session variability, we also conducted a speaker identification experiment using the NTT database. The database consists of sentence data uttered by 35 Japanese speakers (22 males and 13 females) on five sessions over ten months (Aug., Sept., Dec. 1990, Mar., June 1991). In each session, 15 sentences were recorded for each speaker. Ten sentences are common to all speakers and all sessions (A-set), and five sentences

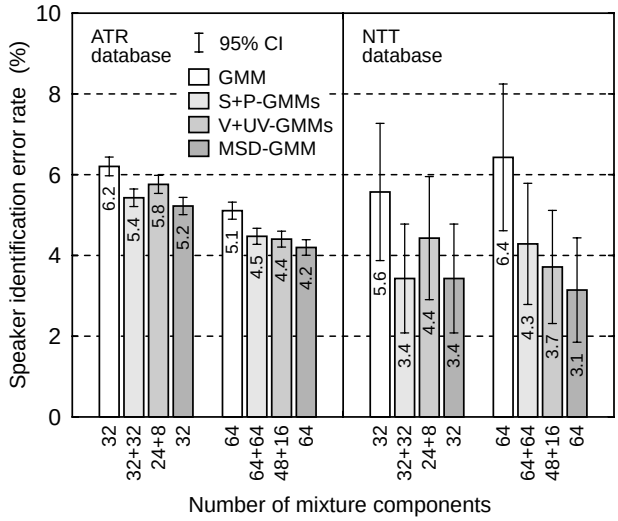

Fig. 4. Comparison of MSD-GMM speaker models with conventional GMM and two-separate GMM speaker models.

are different for each speaker and each session (B-set). The duration of each sentence is approximately four second. We used 15 sentences (A-set + B-set from the first session) per speaker for training, and 20 sentences (B-set from the other four sessions) per speaker for testing. The number of tests was 700 in total.

The speech data were down-sampled to 10 kHz, windowed at a 10-ms frame rate with a 25.6-ms Blackman window, and parameterized into 13 mel-cepstral coefficients using a mel-cepstral estimation technique [9]. The 12 static parameters excluding the zero-th coefficient were used as a spectral feature. Fundamental frequencies (F_0) were estimated at a 10-ms frame rate using the RAPT method [10] with a 7.5-ms correlation window, and $\log F_0$ for the voiced frames and discrete symbols for unvoiced frames were used as a pitch feature. Speakers were modeled by GMMs or multi-stream MSD-GMMs with diagonal covariances.

4.2. Experimental Results

The MSD-GMM speaker identification system was compared with three kinds of conventional systems. Figure 4 shows speaker identification error rates with 95 % confidence intervals (CIs) when using 32 and 64 component speaker models. The left and right halves of the figure correspond to the results for the ATR and NTT databases, respectively. In the figure, "GMM" denotes a conventional GMM speaker model using a spectral feature alone. "S+P-GMMs" represents a speaker model consisting of two GMMs for spectra and pitch. "V+UV-GMMs" is a speaker model consisting of two GMMs for voiced (V) and unvoiced (UV) parts [7] with the optimum numbers of mixture components for the V-GMM and the UV-GMM, i.e., 24 (V)+8 (UV) or 48 (V)+16 (UV), and a linear combination parameter $\alpha = 0.5$ (α is the weight for the likelihood of the UV-GMM). "MSD-GMM" denotes the proposed model based on the multi-stream MSD-GMM.

As shown in the figure, the additional use of pitch information significantly improved the system performance, and the three systems using both spectral and pitch features gave much better performance than the conventional GMM system using a spectral feature alone. Among the three systems, the MSD-GMM system gave the best results, and achieved 16 % and 18 % error reductions (for the ATR database) and 38 % and 51 % error reductions (for the NTT database) over the GMM system when using 32 and 64 mixture models, respectively. It is also noted that the MSD-GMM system requires no combination parameter (such as α) which has to be chosen or tuned heuristically.

5. CONCLUSION

This paper has introduced a new technique for modeling speakers based on MSD-GMM for text-independent speaker identification. The MSD-GMM can model continuous pitch values of voiced frames and discrete symbols representing "unvoiced" in a unified framework. Spectral and pitch features can be jointly modeled by a multi-stream MSD-GMM. We derived the ML estimation formulae for the MSD-GMM parameters and evaluated the MSD-GMM speaker models for text-independent speaker identification tasks. The experimental results demonstrated the high utility of the MSD-GMM speaker model and also proved its robustness against the inter-session variability.

Introduction of stream weights to the multi-stream MSD-GMM and application of this framework to speaker verification systems will be subjects for future works.

6. REFERENCES

[1] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech and Audio Process., 3 (1), 72–83, Jan. 1995.

[2] B.S. Atal, "Automatic speaker recognition based on pitch contours," J. Acoust. Soc. Amer., 52 (6), 1687–1697, 1972.

[3] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," Speech Communication, 5 (2), 183–197, 1986.

[4] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," Proc. ICSLP'96, vol.3, pp.1800–1804, Oct. 1996.

[5] M.K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," Proc. EUROSPEECH'97, vol.3, pp.1391–1394, Sept. 1997.

[6] T. Matsui and S. Furui, "Text-independent speaker recognition using vocal tract and pitch information," Proc. ICSLP'90, vol.1, pp.137–140, Nov. 1990.

[7] K.P. Markov and S. Nakagawa, "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition," J. Acoust. Soc. Jpn., (E), 20 (4), 281–291, July 1999.

[8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proc. ICASSP'99, vol.1, pp.229–232, May 1999.

[9] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. ICASSP'92, vol.1, pp.137–140, Mar. 1992.

[10] D. Talkin, "A robust algorithm for pitch tracking," in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal eds., Elsevier Science, pp.495–518, 1995.