# Speaker Identification under Noisy Environments by Using Harmonic Structure Extraction and Reliable Frame Weighting

*Hiromasa Fujihara†, Tetsuro Kitahara†, Masataka Goto‡, Kazunori Komatani†, Tetsuya Ogata†, and Hiroshi G. Okuno†*

†Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{fujihara,kitahara,komatani,ogata,okuno}@kuis.kyoto-u.ac.jp

‡National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568, Japan
m.goto@aist.go.jp

## Abstract

We present methods for automatic speaker identification in noisy environments. To improve noise robustness of speaker identification, we developed two methods, the *harmonic structure extraction* method and the *reliable frame weighting* method. The harmonic structure extraction method enables the speaker of input speech signals to be identified after environmental noise has been reduced. This method first extracts harmonic components of the speech from the sound mixtures and then resynthesizes a clean speech signal by using a sinusoidal model driven by harmonic components. The reliable frame weighting method then determines how each frame of the resynthesized speech is reliable (i.e. little influenced by environmental noises) by using two Gaussian mixture models for the speech and noise. The speaker can be robustly identified by attaching importance to reliable frames. Experimental results with thirty speakers showed that our method was able to reduce the influences of environmental noise and achieved an error rate of 10.7%, while the error rate for a conventional method was 18.9%.
**Index Terms**: speaker identification, noise robustness, voice extraction, voice reliability, Gaussian mixture model.

## 1. Introduction

Automatic speaker identification is increasing in importance since it can be used in many applications, such as speaker indexing and voiceprint verification. Speaker identification is also useful for human-robot interaction. For example, speaker identification will enable a robot to communicate with any speakers by identifying the voice of a speaker or by detecting speaker's change. In a real environment, however, traditional speaker identification techniques for clean speech voices cannot be used because speech voices are distorted by environmental noises.

To improve the noise robustness of speaker identification, two popular methods have been studied. A spectral subtraction (SS) method [1] reduces a background noise by subtracting the power spectrum of the background noise from an observed power spectrum. A hidden Markov model (HMM) composition method [2, 3] creates noise-added speech HMMs by combining clean speech HMMs with noise HMMs. The noise-added speech HMMs can then be used to identify the speaker. These two methods, however, are not robust to sudden and nonstationary noises because they assume that the power spectrum or HMMs of the background noise are known in advance.

To solve this problem, we propose the *harmonic structure extraction* method and the *reliable frame weighting* method. The harmonic structure extraction method can reduce the influence of environmental noises by extracting the harmonic structure of a speech signal from given audio signals and then resynthesizing the speech signal using a sinusoidal model. This method is robust to unknown nonstationary noises because it does not assume the noise characteristics. On the other hand, the reliable frame weighting method assigns a higher reliability to frames that are less influenced by environmental noises, and uses this reliability as a weight for each frame when identifying the speaker. This method is robust to sudden noises because the contribution from highly distorted frames during speaker identification was reduced. The rest of this paper is organized as follows. The next section describes our method for speaker identification, and Section 3 describes results of our experiments. Finally, Section 4 draws conclusions and points out future directions.

## 2. Speaker Identification Robust To Environmental Noise

Figure 1 shows an overview of our method that identifies a speaker from an input speech signal with background noises. The following describes each step of the method with a focus on the harmonic structure extraction method and the reliable frame weighting method, which are important for the robustness to noises. Although the main idea of these two robust methods is based on our work on automatic singer identification from polyphonic musical audio signals[4], this is the first paper that validates the effectiveness of the idea in identifying a speaker name in speech signals.

### 2.1. Harmonic Structure Extraction

To reduce the environmental noise in a given audio signal, we use a speech resynthesis technique based on harmonic structure. This technique consists of the following three steps:

1. Estimating the fundamental frequency (F0) of the speech using Goto's PreFEst method [5].
2. Extracting the harmonic structure corresponding to the estimated F0 in each frame.
3. Resynthesizing the audio signal (waveform) corresponding to the speech using a sinusoidal synthesis.

We thus obtain a waveform that corresponds to dominant speech.

#### *2.1.1. Fundamental Frequency (F0) Estimation*

We use Goto's PreFEst [5] for estimating the F0 of the speech. The PreFEst method estimates the most predominant F0 in frequency-range-limited sound mixtures. Since the speech tends to have

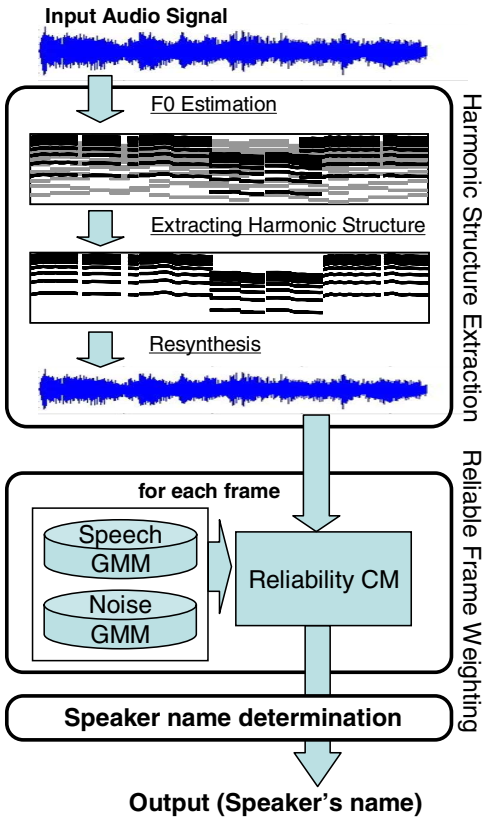



Figure 1: Method overview

the most predominant harmonic structure in middle- and high-frequency regions, we can estimate the F0 of the speech by applying the PreFEst.

A summary of the PreFEst is described below. Hereafter, $x$ is the log-scale frequency denoted in units of cents (a musical-interval measurement) and $(t)$ means time. Given a power spectrum, $\Psi_p^{'(t)}(x)$, we first apply the band-pass filter (BPF) that is designed so that it covers most of the dominant harmonics in typical speech voice. The filtered frequency components can be represented as $BPF(x)\Psi_p^{'(t)}(x)$, where $BPF(x)$ is the BPF's frequency response to the speech. To enable statistical methods to be used, we represent each of the bandpass-filtered frequency components as a probability density function (PDF), called an observed PDF, $p_\Psi^{(t)}(x)$:

$$p_\Psi^{(t)}(x) = \frac{BPF(x)\Psi_p^{\prime(t)}(x)}{\int_{-\infty}^{\infty} BPF(x)\Psi_p^{\prime(t)}(x)dx}. \quad (1)$$

Then, we consider each observed PDF to have been generated from a weighted-mixture model of the tone models for all possible F0s, which is represented as follows:

$$p(x|\theta^{(t)}) = \int_{\mathrm{Fl}}^{\mathrm{Fh}} w^{(t)}(F)p(x|F)dF \quad (2)$$

$$\theta^{(t)} = \{w^{(t)}(F)|\mathrm{Fl} \leq F \leq \mathrm{Fh}\}, \quad (3)$$

where $p(x|F)$ is the PDF of the tone model for each F0, and $\mathrm{Fh}$ and $\mathrm{Fl}$ are defined as the lower and upper limits of the possible (allowable) F0 range, and $w^{(t)}(F)$ is the weight of a tone model that satisfies

$$\int_{\mathrm{Fh}_i}^{\mathrm{Fl}_i} w^{(t)}(F)dF = 1. \quad (4)$$

The tone model represents a typical harmonic structure and indicates where the harmonics of the F0 tend to occur. Then, we estimate $w^{(t)}(F)$ using the EM algorithm and regard it as the F0's PDF. Finally, we obtain the most dominant F0 $\overline{F^{(t)}}$ by the following equation:

$$\overline{F^{(t)}} = \operatorname*{argmax}_F w^{(t)}(F) \quad (5)$$

Since the PreFEst was originally designed to estimate the melody and base lines in musical audio signal, it assumes that an observed spectrum is a mixture of harmonic sounds. Although PreFEst might not perform at its optimum when applied to speech signals in a noisy environment, the estimation result is still useful if used with the reliable frame weighting method for robustness against F0 estimation errors, which is described later.

#### 2.1.2. *Extracting Harmonic Structure*

Based on the estimated F0, we extract the power and phase of the F0 component and harmonic components. For each component, we allow an $|r|$ cent error and extract the peak in the allowed area. The power, $A_l$, phase, $\theta_l$, and frequency, $F_l$ of the $l$ th overtone $(l = 1, \ldots, 20)$ can be represented as

$$F_l = \operatorname*{argmax}_F |S(F)| \quad (l\overline{F} \cdot (1 - 2^{\frac{r}{1200}}) \leq F \leq l\overline{F} \cdot (1 + 2^{\frac{r}{1200}})), \quad (6)$$

$$A_l = |S(F_l)|, \quad (7)$$

$$\theta_l = \arg S(F_l), \quad (8)$$

where $S(F)$ denotes the spectrum and $\overline{F}$ denotes the F0 estimated by the PreFEst. We set $r$ to 20 in our experiments.

Figure 2 shows an example of the F0 estimation and harmonic structure extraction. Figure 2 (a) shows an original spectrum and its envelope, and Figure 2 (b) shows an extracted spectrum and its envelope. These figures show that the spectral envelope of the extracted spectrum (b) correctly represents the formants of speech, when compared with that of the original spectrum (a).

#### 2.1.3. *Resynthesis*

We resynthesize the audio signals of the speech from the extracted harmonic structure by using a sinusoidal model[6]. Resynthesized audio signals are expressed as

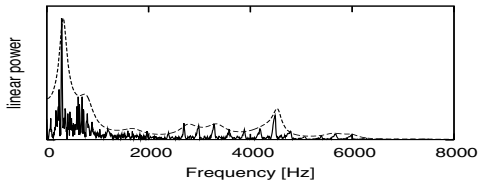
$$s(t) = \sum_{l=1}^{L} A_l \cos(\omega_l t + \theta_l), \quad (9)$$

where $A_l$, $\theta_l$, and $F_l$ represent the power, phase, and frequency of the $l$ th overtone and $t$ is time.

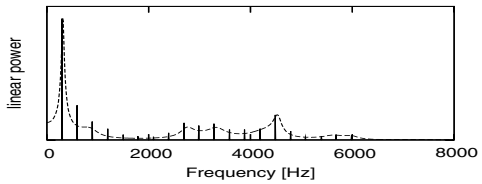
### 2.2. Feature Extraction

We use mel-cepstral coefficients of the LPC spectrum, called linear prediction derived mel-cepstral coefficients (LPMCCs), because we have found that, in the context of singer identification, the LPMCCs can better represent vocal characteristics better than standard mel-frequency cepstral coefficients (MFCCs) [4].

The LPC spectrum is a spectral envelope calculated by linear prediction (LP) analysis. The LP analysis [7][8] is a method that




(a) An original spectrum and its envelope.




(b) An extracted spectrum and its envelope.

Figure 2: Example of F0 estimation and harmonic structure extraction. The envelope of each spectrum is calculated using the linear prediction (LP) analysis.

estimates the transfer function of a vocal tract, assuming that the input audio signal only contains a human voice. In the LP model, when a descrete signal, $s(n)$ is provided, we predict the signal as a linear combination of its previous samples. The predicted $s_W(n)$ value is given by

$$s_W(n) = \sum_{i=1}^{p} \alpha_i s_W(n-i) + g(n), \tag{10}$$

where $p$ represents the order of a predictor, a set of $\alpha_i$ are defined as the linear prediction coefficients (LPCs), and $g(n)$ represents an error in the model. The LPCs are estimated by minimizing the mean squared prediction error of $g(n)$. We used 20th-order LPC in this paper.

In order to calculate the LPMCC, we apply the mel-cepstral analysis to the LPC spectrum. In addition to the role of orthogonalization, the LPMCCs are superior to the LPC in terms of suitability to the human auditory system, which is a benefit of the mel-frequency scale. In this paper, we derive the LPMCC by computing the MFCC from the LPC spectrum for the sake of simplicity of implementation. We set the order of the LPMCC to 15 in this paper.

### 2.3. Reliable Frame Weighting

We introduce two different Gaussian mixture models (GMMs), a speech GMM $\lambda_{\mathrm{V}}$ and a noise GMM $\lambda_{\mathrm{N}}$. The speech GMM $\lambda_{\mathrm{V}}$ is trained using feature vectors extracted from speech signals, and the noise GMM $\lambda_{\mathrm{N}}$ is trained using feature vectors extracted from noise signals. We train each GMM by using the EM algorithm (we use 64-mixture GMMs in our experiments.). Given a feature vector, $\mathbf{x}$, the likelihood that feature vector $\mathbf{x}$ is like speech or a noise can be expressed as $p(\mathbf{x}|\lambda_{\mathrm{V}})$ and $p(\mathbf{x}|\lambda_{\mathrm{N}})$. If the feature vector, $\mathbf{x}$, is less influenced by environmental noise (*i.e.*, more reliable), $p(\mathbf{x}|\lambda_{\mathrm{V}})$ will be higher and $p(\mathbf{x}|\lambda_{\mathrm{N}})$ will be lower. We therefore define reliability of the feature vector, $\mathbf{x}$, as follows:

$$\mathrm{CM}(\mathbf{x}) = \frac{p(\mathbf{x}|\lambda_{\mathrm{V}})}{p(\mathbf{x}|\lambda_{\mathrm{V}}) + p(\mathbf{x}|\lambda_{\mathrm{N}})} w^{(t)}(F), \tag{11}$$

where $w^{(t)}(F)$ is the weight (relative amplitude) of the tone model at the F0 $F$ estimated by the PreFEst in this frame.

Table 1: Dataset

| | |
|---|---|
| Speech data | ASJ-JNAS[9]<br>ATR phonetically-balanced sentences<br>30 speakers (male: 15, female: 15)<br>Training: 10 utterances<br>Evaluation: 40 utterances |
| Noise data | Training: Lobby of the great hall<br>Evaluation: Party venue |

Table 2: Speech analysis

| | |
|---|---|
| Sampling | 16 kHz, 16 bit |
| Window function | Hamming |
| Frame length | 160 ms |
| Frame period | 10 ms |
| Feature vector | 12th order LPMCC<br>based on 20th order LPC analysis |

### 2.4. Speaker Name Determination

The name of the speaker is determined based on the 64-mixture GMMs. We train the GMMs $\lambda_1, \cdots, \lambda_I$ in advance for each registered speaker by using the EM algorithm. Let $\mathbf{X} = \{\mathbf{x}_t | t = 1, \ldots, T\}$ be a time series of feature vectors of input signals and $p(\mathbf{x}|\lambda_i)$ be the likelihood of the GMM for the speaker $i$. Finally, the name of the speaker is eventually determined using the following equation:

$$s = \operatorname*{argmax}_i \frac{1}{T} \sum_{t=1}^{T} \mathrm{CM}(\mathbf{x}_t) \log p(\mathbf{x}_t|\lambda_i) \tag{12}$$

## 3. Experiments

### 3.1. Conditions

To confirm the effectiveness of our methods, we conducted experiments on text-independent speaker identification.

Figure 1 shows the data used in this experiment. We used ASJ-JNAS ATR phonetically-balanced sentences[9] for training and evaluation datasets. The duration of each utterance was approximately 4 seconds. All noise-added data, $x(t)$, were created by mixing clean speech data, $x_S(t)$, with noise data, $x_N(t)$, at 0 dB SNR according to the following equations:

$$x(t) = x_S(t) + k x_N(t), \tag{13}$$

$$k = \sqrt{\frac{S_{\mathrm{POW}}}{N_{\mathrm{POW}}}} 10^{-\frac{\mathrm{SNR}}{10}}, \tag{14}$$

$$S_{\mathrm{POW}} = \frac{1}{T_S} \sum_{t=1}^{T_N} (x_S(t))^2, \tag{15}$$

$$N_{\mathrm{POW}} = \frac{1}{T_N} \sum_{t=1}^{T_N} (x_N(t))^2, \tag{16}$$

where $T_N$ and $T_S$ are the number of samples in the speech and noise sample sequences, respectively. The noise data used in the training were different from those used in the evaluation.

As the training data for the reliable frame weighting, we used 274 utterances from 274 speakers taken from the ASJ-JNAS ATR

Table 3: Experimental results under five different conditions. The "extract." and "weight." denote the harmonic structure extraction method and the reliable frame weighting method, respectively. The "Clean" means that the F0 was estimated from clean speech signals to show the upper limit of the performance. "◯" and "×" mean that the harmonic structure extraction method or the reliable frame weighting method were performed and not performed, respectively.

| Conditions | (i) | (ii) | (iii) | (iv) | (v) |
|---|---|---|---|---|---|
| Noise | with noise (SNR 0 dB) | | | | w/o noise |
| F0 estimation | PreFEst | | - | Clean | - |
| extract. | ◯ | ◯ | × | ◯ | × |
| weight. | ◯ | × | × | × | × |
| Error rate (%) | 10.7 | 26.5 | 18.9 | 5.1 | 1.5 |

phonetically-balanced sentences. Those 274 speakers were not included in the 30 speakers used in the evaluation. We trained speech/noise GMMs using those data after the harmonic structure extracting method had been performed. We did not perform the reliable frame weighting method on the training data of each speaker and used the whole duration of each utterance, while the harmonic structure extraction method was performed.

We conducted experiments under the following five conditions:

**(i)** with extraction using F0s estimated by PreFEst and weighting (proposed),
**(ii)** with extraction using F0s estimated by PreFEst and without weighting,
**(iii)** without both extraction and weighting (baseline),
**(iv)** with extraction using correct F0s estimated from clean speech signal and without weighting (for reference),
**(v)** using clean speech signal for both training and evaluation without both extraction and evaluation (for reference).

Conditions (iv) and (v) were experiments for reference since they used information that could only be obtained from (normally unknown) clean speech signals.

### 3.2. Result and Discussion

Table 3 lists experimental results. The comparison of noised-added speech signals (condition v) with clean speech signals (condition iii) shows that the error rate increased significantly from 1.5% to 18.9%. This increase indicates a bad influence of environmental noises. When our two methods were used together (condition i), the error rate decreased from 18.9% to 10.7%, which is approximately a 43 % reduction in error rate. These results confirmed the improved robustness of our methods to environmental noise.

When we only preformed the harmonic structure extraction method with the F0s estimated from the clean speech (condition iv), the error rate was very low at 5.1%. On the other hand, when we only performed the harmonic structure extraction method with the F0s estimated by PreFEst (condition ii), the error rate was 26.5%, which was even worse than the baseline. This implies that the reliable frame weighting method has high potential, while this also implies that the F0 estimation error has affects the identification error. If we performed the reliable frame weighting method in addition to the harmonic structure extraction method (condition i), the error rate decreased from 26.5% to 10.7%. These results show that the reliable frame weighting method could reduce the negative influence of F0 estimation errors.

## 4. Conclusion

We described the harmonic structure extraction method and the reliable frame weighting method for automatic speaker identification in noisy environments. In our experiments with thirty speakers, we found that our method achieved a 43% decrease in error rate and confirmed the robustness and effectiveness of our methods.

In the future, we plan to extend our method to be able to deal with various severer SNR conditions. Furthermore, we plan to implement this method to robots to achieve robust human-robot speech interaction.

## 5. Acknowledgements

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research (A), No.15200015, and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan). We thank Kazuyoshi Yoshii (Kyoto University) and Takuya Yoshioka (NTT) for their valuable discussions.

## 6. References

[1] Boll, S. F., "Suppression of Acoustic Noise in Speech using Spectral Subtraction," IEEE Transaction on Acoustic, Speech and Signal Processing, Vol. ASSP-27, 113–120, 1979.

[2] Rose, R. C., Hofstetter, E. M., and D. A. Reynolds, "Integrated Models of Signal and Background with Application to Speaker Identification in Noise," IEEE Transaction on Speech and Audio Processing, Vol. 2, no. 2, 245–257, 1994.

[3] Gales, M. J. F., and Young, S. J., "Robust Speech Recognition using Parallel Model Combination," IEEE Transaction on Speech and Audio Processing, Vol. 4, 352–359, 1996.

[4] Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T., and Okuno, H. G., "Singer Identification Based on Accompaniment Sound Reduction and Reliable Frame Selection," In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR2005), 329–336, 2005.

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

[6] Moorer, J. A., "Signal processing aspects of computer music: A survey," In Proceedings of the IEEE, Vol. 65, no. 8, 1108–1137, 1977.

[7] Atal, B. S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," the Journal of the Acoustical Society of America, Vol. 55, no. 6, 1304–1312, 1974.

[8] Shikano, K., "Evaluation of LPC spectral matching measures for phonetic unit recognition," Technical Report CMU-CS-96-108, CMU, Computer Science Department, 1986.

[9] Ito, K., Yamahoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Tokayashi, T., Shikano, K., and Itahashi, S., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," Journal of Acoustic Society Japan (E), Vol.20, No.3, 199–206, 1999.