BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 60, No. 1, 2012
DOI: 10.2478/v10175-012-0011-z

Estimation and tracking of fundamental, 2nd and 3d harmonic frequencies for spectrogram normalization in speech recognition

K. FUJIMOTO[1], N. HAMADA[1], W. KASPRZAK[2*]

[1] Signal Processing Lab., School of Integrated Design Engineering, Keio University, 3-14-1 Hiyoshi, Yokohama 223-8522, Japan
[2] Institute of Control and Computation Eng. Warsaw University of Technology, 15/19 Nowowiejska St., 00-665 Warszawa, Poland

Abstract. A stable and accurate estimation of the fundamental frequency (pitch, $F0$) is an important requirement in speech and music signal analysis, in tasks like automatic speech recognition and extraction of target signal in noisy environment. In this paper, we propose a pitch-related spectrogram normalization scheme to improve the speaker – independency of standard speech features. A very accurate estimation of the fundamental frequency is a must. Hence, we develop a non-parametric recursive estimation method of $F0$ and its 2nd and 3d harmonic frequencies in noisy circumstances. The proposed method is different from typical Kalman and particle filter methods in the way that no particular sum of sinusoidal model is used. Also we tend to estimate F0 and its lower harmonics by using novel likelihood function. Through experiments under various noise levels, the proposed method is proved to be more accurate than other conventional methods. The spectrogram normalization scheme makes a mapping of real harmonic structure to a normalized structure. Results obtained for voiced phonemes show an increase in stability of the standard speech features – the average within-phoneme distance of the MFCC features for voiced phonemes can be decreased by several percent.

Key words: automatic speech recognition, spectrogram analysis, particle filter, pitch estimation.

1. Introduction

The *fundamental* frequency ($F0$) plays an important role in human speech generation [1, 2]. But the information about $F0$ has only rarely been used in the past to improve automatic speech recognition systems [3, 4]. Instead, initially the interest concentrated here on *formants*, i.e. the resonances of the vocal tract, which appear on spectrograms as regions of high energy [5]. They efficiently describe essential aspects of speech using a very limited set of parameters. Variability of formant locations for different speakers are explained by different lengths of the vocal tract. Formant estimation corresponds to the LPC (linear predictive coding) approach to speech representation. Since the mid-1980s the most popular representation in speech recognition has become the mel-frequency cepstral coefficients (MFCC) [6]. A more recent approach is called LPCC (linear predictive cepstral coefficients) [6] and it is related to the two-tube and three-tube models of the vocal tract [5].

Here, we want to utilize the advantage of detecting harmonic frequencies of the pitch for a spectrogram normalization step, that improves the speaker-independence of MFCC features. Thus, a reliable estimation of the pitch and a pitch-related normalization of the speech signal is the scope of this paper.

A reliable pitch detection (which is $F0$) is generally a main step in speech and music signal analysis, dedicated to the extraction of target signal in noisy environment. Most $F0$ estimation methods are based on autocorrelation, the amplitude magnitude difference function (AMDF), the cepstrum analysis or linear prediction (LP or PARCOR) [7–11]. They are very effective for a noiseless target speech. For real environments, robust techniques, less sensitive to background noise and reverberation, have been developed, like comb-filter approach [12], instantaneous amplitude (IA) and frequency (IF) approaches [13–15]. Another type of $F0$ estimation approach is to use the recursive state estimation scheme such as Kalman filter and particle filter [16–19].

In this paper we propose a method for a stable and accurate estimation of the fundamental frequency (F), and consider it as an important requirement for a speech spectrogram normalization approach. The proposed method makes a non-parametric recursive estimation of $F0$ and its 2nd and 3rd harmonic frequencies in noisy circumstances. Therefore, our method is different from standard Kalman and particle filter methods in the way that no particular sum of a sinusoidal model is used, like proposed in [20]. Also we tend to estimate $F0$ and its lower harmonics by using a novel likelihood function. Its goal is to estimate $F0$ by reflecting the spectrum at $2F0$. Then, the individual likelihood functions associated with estimating 2nd and 3rd harmonics are proposed.

The common approach to make speech features speaker-independent is to perform a linear or piecewise linear warping of the frequency axis. The warping function for an individual speaker is estimated by a maximum-likelihood (ML) approach [21–22]. Obviously, this requires a large training material to be collected for a speaker in advance. In contrast to ML-based approaches there exist very few on-line approaches to speaker-independent feature normalization. They usually explore the locations of main spectral formants [23–24] as it is known that they correspond to specific features of the speaker's vocal tract.

*e-mail: W.Kasprzak@elka.pw.edu.pl

In the frequency domain voiced speech signal frames have spectral peaks at or near integer multiples of the fundamental frequency. Most simply the pitch modification can be achieved by multiplying the frequency representation by a sinusoidal wave [25]. However, this will result in the strengthening of side lobs and noise – making the voice very unnatural. Another approach is to make formant-corrected pitch shift [3]. The spectral envelope must be preserved in the modification stage so as to preserve the formant structure of the vocal tract. Thus the amplitudes at the new harmonic frequencies are achieved by sampling the spectral envelope of the original signal at its harmonic frequencies, whereas at remaining – non-harmonic frequencies the amplitudes being interpolations of original non-harmonic frequencies are set.

The paper is organized into 4 more sections. Section 2 introduces the spectrogram normalization approach and the recursive estimation method, that is based on a particle filter. In Sec. 3, the new estimation and tracking method is proposed and especially, likelihood functions are introduced. Several experimental results of $F0$ estimation and spectrogram normalization are described in Sec. 4. Through experiments under various SNR, the proposed $F0$ estimation method is proved to be more accurate than conventional methods [26-30] Finally conclusions are drawn in Sec. 5.

2. Problem

Let us explain the motivation for spectrogram normalization on base of the following example. In Fig. 1 there are shown two different Fourier magnitude distributions in a single signal frame, that contains the same vowel /e/. Two observations can be made. It is clearly visible that the magnitude peaks are shifted w.r.t. each other. The "green" voice has a higher $F0$ then the "blue" voice, but a lower formant $F1$, again a higher formant $F2$, and in turn a lower formant $F3$. The second observation is, that the "green" curve is smoother than the other one. This is due to a lower number of harmonic frequencies contained in given interval of frequencies (because of higher $F0$). We want to generalize these observations and to propose a normalization scheme in which: 1) the fundamental frequency will be set to a default average frequency, 2) at default harmonic frequencies the magnitudes of Fourier coefficients will be approximated by magnitudes of "nearest" real harmonic frequencies, and 3) the approximation weights express the required orientation of the shift.

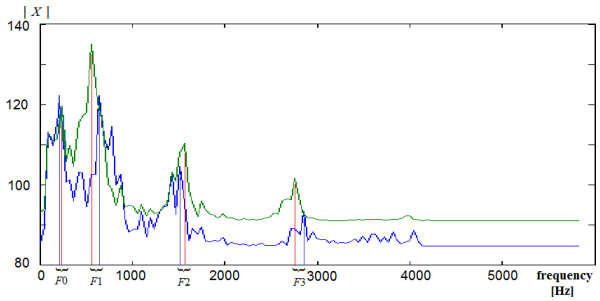


Fig. 1. Example of two distributions of the magnitude of Fourier coefficients for vowel /e/ The differences in formant locations are marked by red and blue lines

2.1. *F0*-based spectrogram normalization. The spectrogram normalization procedure consists of three steps:

1) On-line estimation of current $F0$.

The non-parametric approach to $F0$ estimation, based on particle filtering, is described in Sec. 3.

2) Determining current phoneme type.

At first, we need to determine the frames containing speech. A simple voice activity detector (VAD) is applied, that uses an energy-based adaptive threshold (see Subsec. 2.4) in the decision rule. The discrimination between voiced and unvoiced phonemes can be done in different way, in time or frequency domain. Our approach is to use conditions posed onto auxiliary speech features: the low-pass ratio and the maximum autocorrelation factor, computed for a single signal frame.

The low-pass ratio is defined as the relation of the sum of magnitudes of Fourier coefficients X_k, given in the low-frequency band of 60–1000 Hz and in the entire band of 60–6000 Hz:

$$\rho_{LP}^{(m)} = \frac{\sum\limits_{k\geq 60/f_s}^{k\leq 1000/f_s} |X_k^{(m)}|}{\sum\limits_{k\geq 60/f_s}^{k\leq 6000/f_s} |X_k^{(m)}|}. \tag{1}$$

The maximum of a normalized autocorrelation coefficient for a frame of signal samples $[x_\tau, \ldots, x_{\tau+N-1}]$ starting at sample τ:

$$\rho_{\max r}^{(\tau)} = \max_{k=2,\ldots,N/4} r_k^{(\tau)} = \frac{\sum\limits_{n=\tau}^{\tau+N-k-1} x_n x_{n+k}}{|[x_n]||[x_{n+k}]|}. \tag{2}$$

The voiced/unvoiced discrimination rule:

$$\zeta(\rho_{LP}^{(m)}, \rho_{\max r}^{(\tau)}) = \begin{cases} 1 & (\rho_{LP}^{(m)} \geq 0.4) \wedge (\rho_{\max r}^{(\tau)} \geq 0.8) \\ & \text{or } (\rho_{LP}^{(m)} + \rho_{\max r}^{(\tau)}) \geq 1.3 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

3) Normalization of the FC coefficients.

The mapping of the magnitude of Fourier coefficients (FC) from the original vector spanned over the harmonics of the pitch frequency, $[F_0, 2F_0, 3F_0, \ldots]$, to the new vector spanned over the harmonics of the normalized pitch frequency, $[F_N, 2F_N, 3F_N, \ldots]$, is described as follows (for illustration see Fig. 2):

a) The original vector of Fourier coefficient magnitudes is given (e.g. the frequency domain is sampled at multiple frequencies of 40 Hz). Assume that the current pitch frequency is: $F_0 = 162$ Hz. From the envelope line it is visible that the first two formants are located near the 320 and 1120 Hz (in fact according to our pitch assumption correct values are 324 and 1134 Hz) Fig. 2a.

b) At first in the output distribution, the magnitudes of Fourier coefficients are approximated at multiple frequencies of the normalized frequency (e.g. $F_N = 120$ Hz) – computed from the original magnitude envelope as weighted sums of original FC located at two nearest harmonic frequencies of pitch F_0 (Fig. 2b).

c) At second, the remaining Fourier coefficients in the output distribution are set to the average value of original FC magnitudes for frequencies located between harmonic frequencies of the original pitch (Fig. 2c).

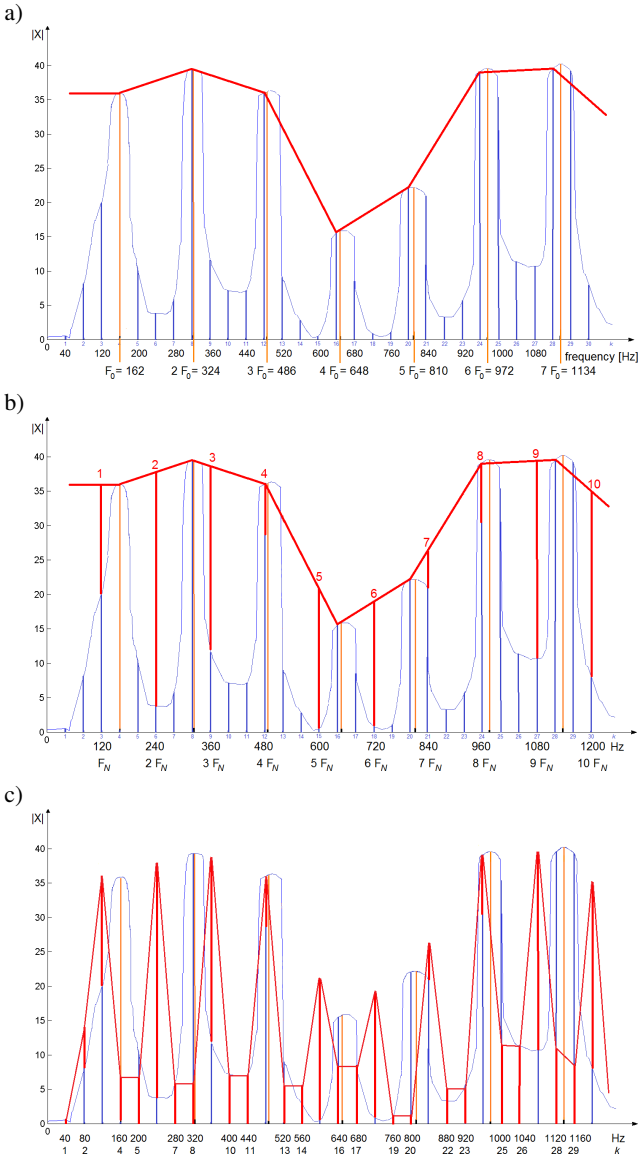


Fig. 2. Illustration of the spectrogram normalization procedure, induced by the change of pitch frequency from 162 Hz to 120 Hz: (a) Original distribution of Fourier coefficients (FC) (bars located at multiple of 40 Hz frequencies), the assumed original pitch frequency at 162 Hz and its harmonics, and the magnitude envelope line; (b) Approximation of FC magnitudes at harmonic frequencies of the normalized pitch of 120 Hz taken from the magnitude envelope of the original distribution; (c) Approximation of FC magnitudes at frequencies located between new harmonic frequencies, and the final magnitude envelope

2.2. The *F*0 estimation problem. The robustness of recursive estimation schemes against non-stationary noise is due to the fact that the estimated states in previous frames are recursively used for the state estimation in current frame of the signal. Usually a parametric model is applied, which assumes a sum of lower harmonic components plus additional random noise signal. The parameters of such model are $F0$ and its IA, the IAs of several harmonics, and the number of harmonics. The state transition and observation equations are assumed, and then the extended Kalman filter algorithm or Bayesian parameter estimation scheme is applied for recursive estimation of the time-varying parameters. Recently, a particle filter has been proposed for recursively estimating local peaks of speech spectrum in noisy environment [18]. The method improves the robustness of conventional peak-picking methods [29] by novel two-step particle filter approach. The first step utilizes likelihood of peaks using spectral envelope of the cepstrum, and the second step determines peaks from the frequency band taking highest peak presence probability.

Time series filtering. Modeling of dynamic system and observation is an important concept in time series filtering using regressive algorithm. If the state transition is modeled by Markov property, observed vectors $\mathbf{y}_t$ and state space vectors $\mathbf{x}_t$ can be expressed as follows.

$$\textbf{System model}: \quad \mathbf{x}_{t+1} = \mathbf{f}_t(\mathbf{x}_t) + \mathbf{w}_t, \tag{4}$$

$$\textbf{Observation model}: \quad \mathbf{y}_t = \mathbf{h}_t(\mathbf{x}_t) + \mathbf{v}_t. \tag{5}$$

$\mathbf{f}_t(\mathbf{x}_t)$ is the state transition function from time $t$to t+1, $\mathbf{h}_t(\mathbf{x}_t)$ is the observation function which expresses the relation between states and observed values, $\mathbf{w}_t$, $\mathbf{v}_t$ are white noises.

The state estimation problem results in the estimation of posterior probability distribution of $\mathbf{x}_t$ which is expressed as $p(\mathbf{x}_t|\mathbf{Y}_t)$ with the set of observation series $\mathbf{Y}_t = \{\mathbf{y}_1 \dots \mathbf{y}_t\}$. On the contrary, it is difficult to estimate $p(\mathbf{x}_t|\mathbf{Y}_t)$ directly from $\mathbf{Y}_t$. Therefore, this problem is solved by converting it to sequential estimation using Bayes' theorem and Markov property of the state space. Based on Bayes' theorem, $p(\mathbf{x}_t|\mathbf{Y}_t)$ can be computed by the multiplication of likelihood and prior probability $p(\mathbf{x}_t|\mathbf{Y}_{t-1})$ as follows:

$$p(\mathbf{x}_t|\mathbf{Y}_t) = \frac{p(\mathbf{Y}_t|\mathbf{x}_t)}{p(\mathbf{Y}_t|\mathbf{Y}_{t-1})} p(\mathbf{x}_t|\mathbf{Y}_{t-1}). \tag{6}$$

Here, the likelihood $p(\mathbf{Y}_t|\mathbf{x}_t)$ expresses the probability to observe $\mathbf{Y}_t$ in the certain state. $p(\mathbf{Y}_t|\mathbf{Y}_{t-1})$ is the normalization term which makes $(\int p(\mathbf{x}_t|\mathbf{Y}_t)d\mathbf{x}_t\ )$ equal to 1 and has no relation with $\mathbf{x}_t$. $p(\mathbf{x}_t|\mathbf{Y}_{t-1})$ is the prior probability of $\mathbf{x}_t$ at the time t, and is given by

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{Y}_{t-1}) &= \int p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{Y}_{t-1}) d\mathbf{x}_{t-1} \\ &= \int p(\mathbf{x}_t|\mathbf{x}_{t-1}\mathbf{Y}_{t-1}) p(\mathbf{x}_{t1}|\mathbf{Y}_{t-1}) d\mathbf{x}_{t-1} \\ &= \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1}) d\mathbf{x}_{t-1}, \end{aligned} \tag{7}$$

based on Markov property. Here, $p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1})$ is the posterior probability at the time $t-1$, and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the state transition probability from time $t-1$ to t, and it is given by Eq. (4). As just described, the filtering task is to estimate posterior probability $p(\mathbf{x}_t|\mathbf{Y}_t)$ at each time. It is performed by

the prediction of the state at time $t$using observed value up to time $t-1$ and state transition probability.

Particle filter. Except for its extended version, the Kalman filter poses restrictions on both observing and system models such as being linear and Gaussian model. In contrast, these restrictions are not imposed on the particle filter (PF). Therefore, PF is a valid approach to estimate nonstationary, nonlinear, and nonGaussian state space model. In PF the non-Gaussian distribution is approximated by particles: $\mathbf{S}_t = [\mathbf{s}_t^{(1)}, \ldots, \mathbf{s}_t^{(N)}]$. Therefore, the pdf-s $p(\mathbf{x}_t|\mathbf{Y}_t)$ and $p(\mathbf{x}_t|\mathbf{Y}_{t-1})$, that usually appear in the time series filtering are approximated by using sufficiently many particles. When N particles exist, *prior* probability and *posterior* probability are approximated by:

$$\text{Prior:} \quad p(\mathbf{x}_t|\mathbf{Y}_{t-1}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(\mathbf{x}_t - \mathbf{s}_{t|t-1}^{(i)}), \tag{8}$$

$$\text{Posterior:} \quad p(\mathbf{x}_t|\mathbf{Y}_t) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(\mathbf{x}_t - \mathbf{s}_{t|t}^{(i)}) \tag{9}$$

Here we have two particle sets $\mathbf{S}_{t|t-1}$=$\{\mathbf{s}_{t|t-1}^{(1)}, \ldots \mathbf{s}_{t|t-1}^{(N)}\}$ and $\mathbf{S}_{t|t}$=$\{\mathbf{s}_{t|t}^{(1)}, \ldots \mathbf{s}_{t|t}^{(N)}\}$. $\mathbf{S}_{t|t}$ is the posterior distribution of particles which are wiped out or copied from the prior (predicted) distribution $\mathbf{S}_{t|t-1}$ depending on observation at time t, that results in weighting (likelihood) of particles

The essence of PF is estimating posterior distribution $p(\mathbf{x}_t|\mathbf{Y}_t)$ by using posterior particle set $\mathbf{S}_{t|t}$. The $\mathbf{S}_{t|t}$ is generated as a weighted version of prior particles $\mathbf{S}_{t|t-1}$ To realize this, we generate both sets of prior particles $\mathbf{S}_{t|t-1}$ and posterior particles $\mathbf{S}_{t|t}$ at certain time t based on the following steps known as the sampling importance resampling filter [32].

1) Initialization

Set the initial arrangement of particles $\mathbf{S}_{t-1}0$. Set $t = 1$.

2) Prediction of next prior distribution of particles

Add the random noise to each particle $\mathbf{s}_{t1|t-1}^{(i)}$ $(i = 1, \ldots, N)$ and scatter them according to the system model, $p(\mathbf{x}_t|\mathbf{x}_{t-1}$=$\mathbf{s}_{t-1|t-1}^{(i)})$, to produce the prior distribution of particles as follows:

[2-1] Generate a series of random system noise $\mathbf{w}_{t-1}^{(i)}$

[2-2] Generate prior predicted distribution of particles by transition process of particles at time $t-1$ based on system model Eq. (4):

$$\mathbf{S}_{t|t-1} = \{\boldsymbol{f}_{t-1}(\mathbf{s}_{t-1|t-1}^{(i)}\mathbf{w}_{t-1}^{(i)}) \quad |i = 1, \ldots, N\}. \tag{10}$$

3) Weighting of particles due to observation

Calculate the current likelihood $p(\mathbf{y}_t|\mathbf{x}_t = \mathbf{s}_{t|t-1}^{(i)})$ based on the observation model, Eq. (5), then update the weights $\pi_{t|t-1}^{(i)}$ of i-th prior predicted particle as

$$\pi_t^{(i)} = \frac{p(\mathbf{y}_t|\mathbf{x}_t = \mathbf{s}_{t|t-1}^{(i)})}{\sum_{i=1}^{N} p(\mathbf{y}_t|\mathbf{x}_t = \mathbf{s}_{t|t-1}^{(i)})} \tag{11}$$

to define the set $\Pi_t = \left\{\pi_t^{(i)}|i = 1, ..., N\right\}$

4) Update – posterior distribution of particles

[4-1] Resampling

The set of particles and their weights, $\left\{\mathbf{s}_{t|t-1}^{(i)}, \pi_i^{(t)}\right\}$, is mapped into a consistent set with uniform weights $\left\{\mathbf{s}_{t|t}^{(i)}, N^{-1}\right\}$.

[4-2] The updated particle set approximates the posterior probability:

$$p(\mathbf{x}_t|\mathbf{Y}_t) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(\mathbf{x}_t - \mathbf{s}_{t|t}^{(i)}) \tag{12}$$

[4-3] The posterior particles t also allow to estimate the state variable as:

$$\widehat{\mathbf{x}}_t = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_{t|t}^{(i)} \tag{13}$$

5) Termination test

If $(t < T)$ then set $t = t+1$ and go back to the prediction step 2) else stop.

3. Proposed $F0$ estimation approach

Figure 3 shows the flow of the $F0$ estimation process performed with the help of a particle filter. In particular we need to explain the meaning of particles, the measurement process, the weighting (likelihood) setting for harmonic frequencies and the update of particles

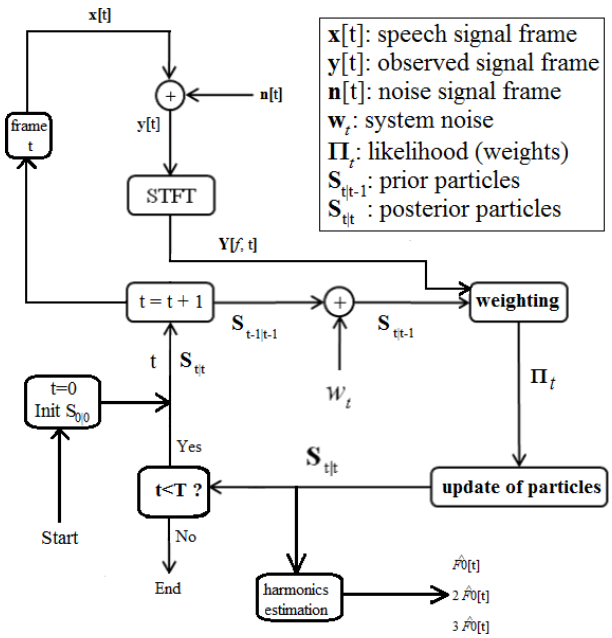


Fig. 3. Flow of the $F0$, $2F0$, $3F0$ estimation

3.1. Problem representation. Consider a discrete-time signal of finite length:

$$y[k] = x[k] + n[k], \tag{14}$$

where $x[k]$ is the source speech signal, $n[k]$ is noise, and k is the discrete time index. Applying a windowed L-point short

time Fourier transform (STFT) we get a matrix of observations in the frequency space – the spectrogram:

$$Y[f,t] = X[f,t] + N[f,t], \tag{15}$$

where $t = 1, 2, \ldots, T$; is the integer index of time frame, and f indicates the index of frequency bin, where $f = 0, 1, \ldots, L/2$.

A set of particles $\mathbf{S}_{t|t}$ represents the likelihood of power distribution in the frequency domain. It is initialized according to local power maxima and then, due to different measurement steps for $F0$, the 2nd and 3d harmonic frequencies, there are three set of particles $\mathbf{S}_{t|t}$ iteratively obtained, for every $t = 1, \ldots, T$. Current estimation of $F0$, the 2nd and 3d harmonic frequencies is done in sequence, based on the information of corresponding particle distribution $\mathbf{S}_{t|t}$.

In the next iteration the particles are resampled by random noise $\mathbf{w}_t$ and due to next measurement step, they are weighted again. This allows again a resampling that concentrates on particles expressing the currently expected harmonic structure of the speech signal. Figure 4 illustrates the resampling steps done for samples aimed to represent the fundamental frequency $F0$.

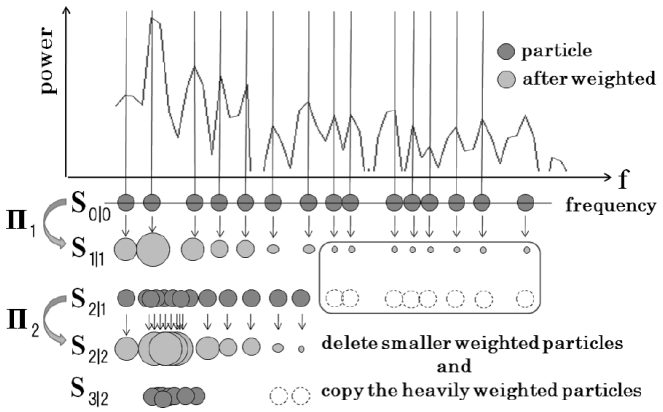


Fig. 4. Resampling of particles (during the prediction and update steps)

3.2. Likelihood setting (weighting). The most significant factor in the proposed estimation is likelihood setting. The likelihood is determined based on power spectrum and its harmonic structure.

First, our attention is focused on the local maximum power spectrum at the actual $F0$. However, the local maximum power property does not always ensure the real $F0$, because observed power spectrum is sensitive to environmental noise. In addition, power level of 2nd harmonic may happen to be larger than the power level at $F0$. To avoid these problematic cases, we need to incorporate harmonic structure. That is, we may find local maximum power at some integer multiplied frequency of $F0$. This fact is utilized for next likelihood setting.

(1) Likelihood for $F0$ determination

The following likelihood is defined as the weighting term of the i-th prior particle $s^{(i)}_{t|t-1}$ (in $F0$ estimation we use now the notation $\pi^{(i)}_{F_0}[t]$ for weight $\pi^{(i)}_t$):

$$\pi^{(i)}_{F_0}[t] = \alpha\xi^{(i)}_t + (1-\alpha)\psi^{(i)}_t, \tag{16}$$

$$\xi^{(i)}_t = \frac{\left(\left\|\mathbf{Y}(s^{(i)}_{t|t-1}, t)\right\| + \left\|\mathbf{Y}(2s^{(i)}_{t|t-1}, t)\right\|\right)^2}{\sqrt{\sum_{i=1}^{N}\left(\left\|\mathbf{Y}(s^{(i)}_{t|t-1}, t)\right\| + \left\|\mathbf{Y}(2s^{(i)}_{t|t-1}, t)\right\|\right)^2}}. \tag{17}$$

The term $\xi^{(i)}_t$ represents the normalized mixed power at the points of i-th particle located at frequency F and at $2F0$. Due to harmonic structure of voiced speech, when i-th particle exists at actual $F0$, the first term of the right hand side of Eq. (16) takes large value. Because it reflects the power at twice frequency of particle $s^{(i)}_{t|t-1}$, this likelihood prevents from misestimating 2nd harmonics as actual $F0$. The second term $\psi^{(i)}_t$ of the right hand side of Eq. (16) is binary number which indicates whether the estimated point meets the local peaks properly. Namely, we set $\psi^{(i)}_t = 1$ if both of the following peak conditions are satisfied:

$$\begin{aligned}\left\|\mathbf{Y}(s^{(i)}_{t|t-1}, t)\right\| &> \left\|\mathbf{Y}(s^{(i)}_{t|t-1} \pm \varepsilon, t)\right\|,\\ \left\|\mathbf{Y}(2s^{(i)}_{t|t-1}, t)\right\| &> \left\|\mathbf{Y}(2s^{(i)}_{t|t-1} \pm \varepsilon, t)\right\|.\end{aligned} \tag{18}$$

Otherwise, $\psi^{(i)}_t = 0$. We set $\alpha = 0.6$ experimentally, and ε is the frequency unit which corresponds to one frequency bin width.

(2) Likelihood for 2nd and 3d harmonics

Next step is to establish the likelihood for harmonics. We define the likelihood for k-th harmonics ($k = 2, 3$) by

$$\begin{aligned}\pi^{(i)}_{harm}[t] = \beta&\frac{\left(\left\|\mathbf{Y}(s^{(i)}_{t|t-1}, t)\right\|\right)^2}{\sqrt{\sum_{i=1}^{N}\left(\left\|\mathbf{Y}(s^{(i)}_{t|t-1}, t)\right\|\right)^2}}\\ &+ \frac{1-\beta}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(s^{(i)}_{t|t-1} - k\widehat{F}_0)^2}{2\sigma^2}\right).\end{aligned} \tag{19}$$

As the order of harmonics increases, its power tends to decrease. From this observation, it is not proper to take the power value of harmonic itself as the likelihood. In the proposed method, the elements selected as the likelihood are the power of the points where particles exist. For the former, we set the first term of the right side of Eq. (19) which takes higher value when i-th particle exists at actual harmonics. The second term of the right side means the current probability of harmonics. It is based on normal distribution whose mean is $k(= 2, 3)$ times the value of the estimated fundamental frequency $F0$, and the standard deviation σ is determined experimentally. In later experiment we set it the value which corresponds to width of 8 frequency bins.

3.3. *F*0, 2nd, 3rd harmonics estimation. Due to managing nonlinear, nonGaussian model in PF, the estimated $F0$ is

obtained by the average of particles, namely the arithmetic weighted mean of the posterior N particles given by:

$$\widehat{F}_0[t] = \frac{1}{N}\sum_{i=1}^{N} s_{t|t}^{(i)} = \frac{1}{N}\sum_{i=1}^{N} \pi_{F_0}^{(i)} s_{t|t-1}^{(i)} \qquad (20)$$

is used based on Eqs. (8), (9). Same weighted average computation is adopted for estimating the 2nd and 3d harmonic frequencies by using related particles.

3.4. Voice activity detector. One problem in the proposed method as well as in other recursive estimation schemes is that $F0$ is intended to estimate even in unvoiced and silent frames, what will result in false results. To avoid this, we need to use a VAD (voice activity detector) and to provide a voiced/unvoiced discrimination rule We adopt a simple VAD criterion: if the power at estimated frequency each particle exists does not attain a specific threshold, namely the condition

$$|Y(s_{t|t}^{(i)}, t)|^2 < \chi \ \mathrm{Max}|Y|^2 \qquad (i = 1, \ldots, N) \qquad (21)$$

is satisfied, then no update of samples in current frame is performed and the estimated value is also not adopted as a $F0$. $\mathrm{Max}|Y|^2$ is the maximum value of power spectrum of speech across the time interval. We set $\chi = 0.03$ in the experiments.

As explained in Subsec. 2.1, the unvoiced/voiced discrimination rule is based on two speech parameters computed for every frame: the low-pass ratio and the maximum value of auto-correlation coefficient. We run the estimation for both voiced and unvoiced signal frames but vary the parameter α for them.

4. Experiments

At first, the proposed method of $F0$ estimation is applied to speech signals with two types of additive noise. The first one is white noise, and the other is a sample of real environment noise.

At second, the correct estimation of $F0$ is used to normalize the spectrogram and to compute MFCC features. These features of selected signal frames are evaluated at the end.

4.1. *F0* estimation under white noise. Evaluation criteria. In order to compare quantitatively these obtained results with the results by several conventional methods, we defined two measures: gross error rate and fine error rate.

Gross error rate (GER) is the ratio of the number of time frames giving "incorrect" values to the total number of frames. Value of $F0$ is called "incorrect" if it falls outside $\pm 10\%$ of the actual $F0$ value.

Fine error rate (FER) is the ratio of the number of time frames giving "correct" values to the total number of frames. Value of $F0$ is called "correct" if it falls inside $\pm 5\%$ of the actual $F0$ value. We can see the robustness of the estimation against noise from the GER, and the accuracy of estimation from the FER.

Robustness of the proposed method against white noise is confirmed as follows. Four SNR cases, such as 20 dB, 10 dB, 5 dB, and 0 dB are examined. Before the error evaluation is performed actual $F0$ and 2nd, 3d harmonics frequencies are estimated. For comparing the proposed method with other methods, most rigorous way is to use database with EGG data as reported in [21]. On the other hand, we determine the actual $F0$ and harmonic frequencies manually by observing the amplitude values in time-frequency plane of the original (highest SNR) speech signal manually. The accuracy in this paper means how the estimated frequencies are apart from these manually determined results.

Comparative results. The proposed method is compared with following four conventional methods:

1. The auto-correlation method [26, 27] (Auto-correlation method for period determination),
2. The amplitude magnitude difference function (AMDF) [28] (Time-lag search for globally maximizing the magnitude of difference function),
3. The linear prediction (LP) and residual signal method (PARCOR) [29] (Using the LP source-filter model for detecting glottal characteristics),
4. The cepstrum-based method [30] (Cepstrum or homomorphic model for detecting glottal characteristics).

Since these conventional methods usually estimate solely $F0$, then, the 2nd and 3d harmonics estimations are set to be simply the double and triple values of $F0$. There exist other methods such as [33] etc., The method [33] for instance was compared with AMDF in [28]. In addition, [34] reported the comparative evaluation for several estimation methods.

Additionally, in order to prevent nonessential miss-estimation when we apply the conventional methods, two thresholds are set. They prevent from miss-estimations that are less than the half and more than twice of actual $F0$. However, these miss-estimations do not occur in the proposed method. Experiments were performed for four (2 males & 2 females) speech signals from the *Acoustic Society of Japan (ASJ) continuous speech corpus for research*. We added white Gaussian noise to the sources.

The results of harmonic frequencies estimation are shown in Fig. 5. As observed in these results, the proposed method gives accurate both $F0$ and 2nd harmonic estimation even in low SNR cases. On the other hand, 3d harmonic frequency is occasionally misestimated in lower SNR case. It is due to the decrease of power at higher harmonic frequency.

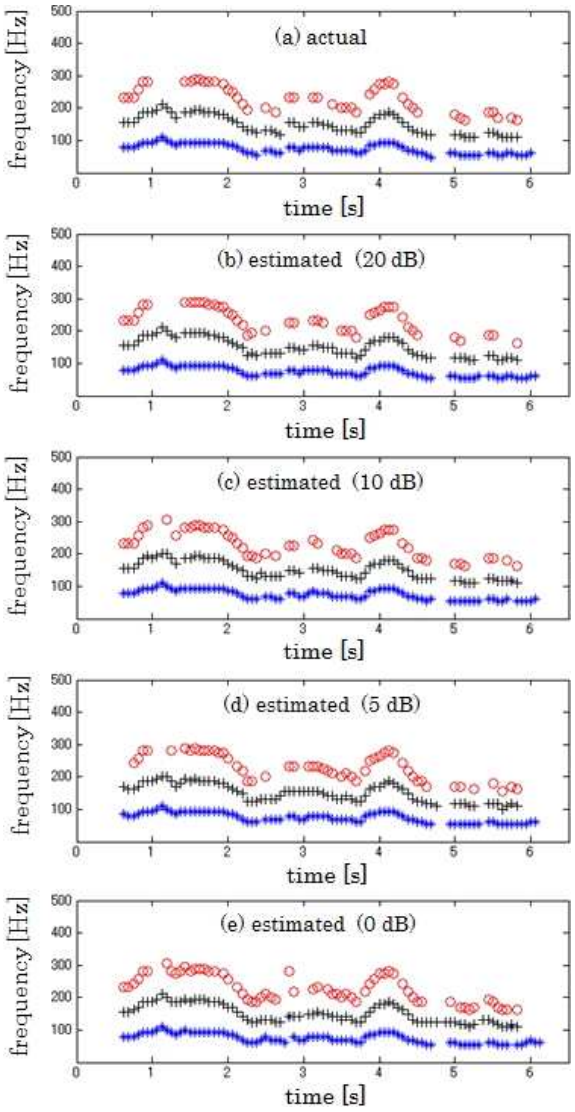


Fig. 5. (a) Actual contour, (b)–(e) estimation results in each SNR, (b) 20 dB, (c) 10 dB, (d) 5 dB, (e) 0 dB. Estimated $F0$, 2nd and 3d harmonics are marked by "$*$", "+", "o" respectively

This implies the advantage of the proposed method. Figure 6 shows the estimation results of each method under the 10 dB white noise. In addition, GER results in Fig. 7 show that the proposed method can estimate $F0$, $2F0$, $3F0$ robustly against noise. PARCOR method can also estimate robustly, while, we can say that the proposed method gives more accurate estimations than other methods associated with FER.

In lower SNR condition, the original shape of speech spectrum would be more or less deformed; therefore, the conventional methods based on the periodicity of voiced signals does not cope with these cases. Meanwhile, the proposed method comparatively gives accurate results. It is resulted by introducing proper likelihood which is able to evaluate how the estimated $F0$ is appropriate. In addition, it would satisfy temporal continuity.

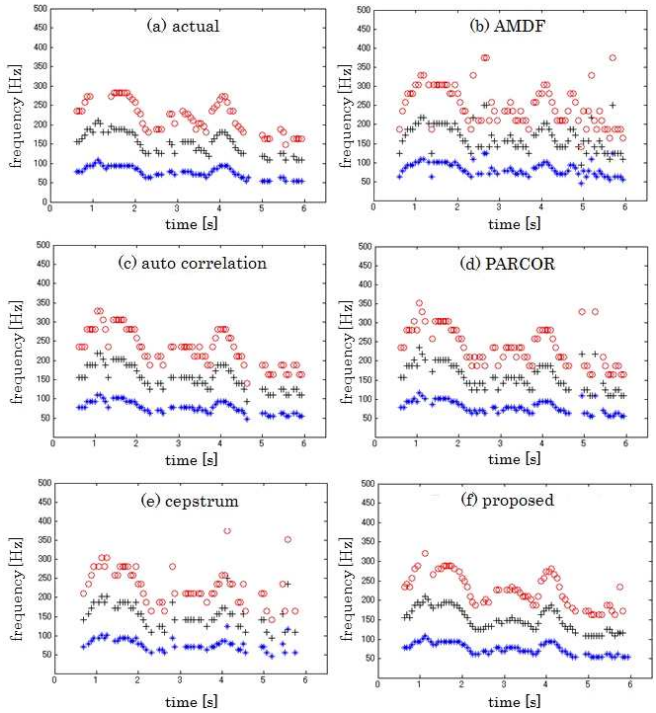


Fig. 6. Actual harmonic profile and estimated results obtained by improved conventional and the proposed methods (SNR=10 dB white noise case)

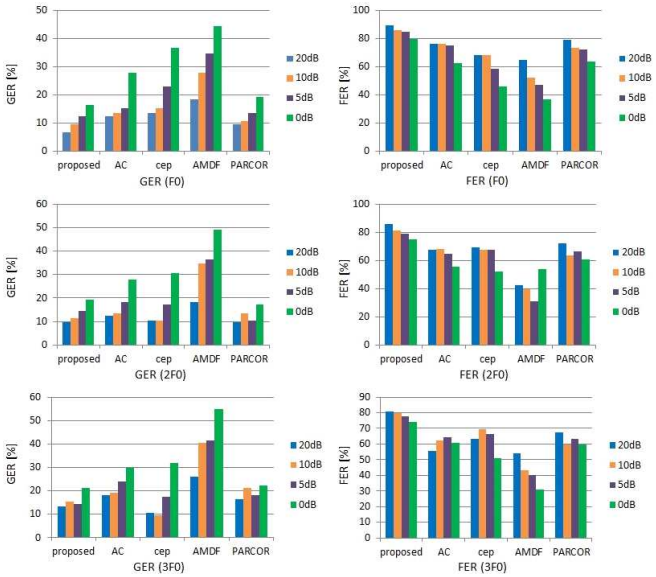
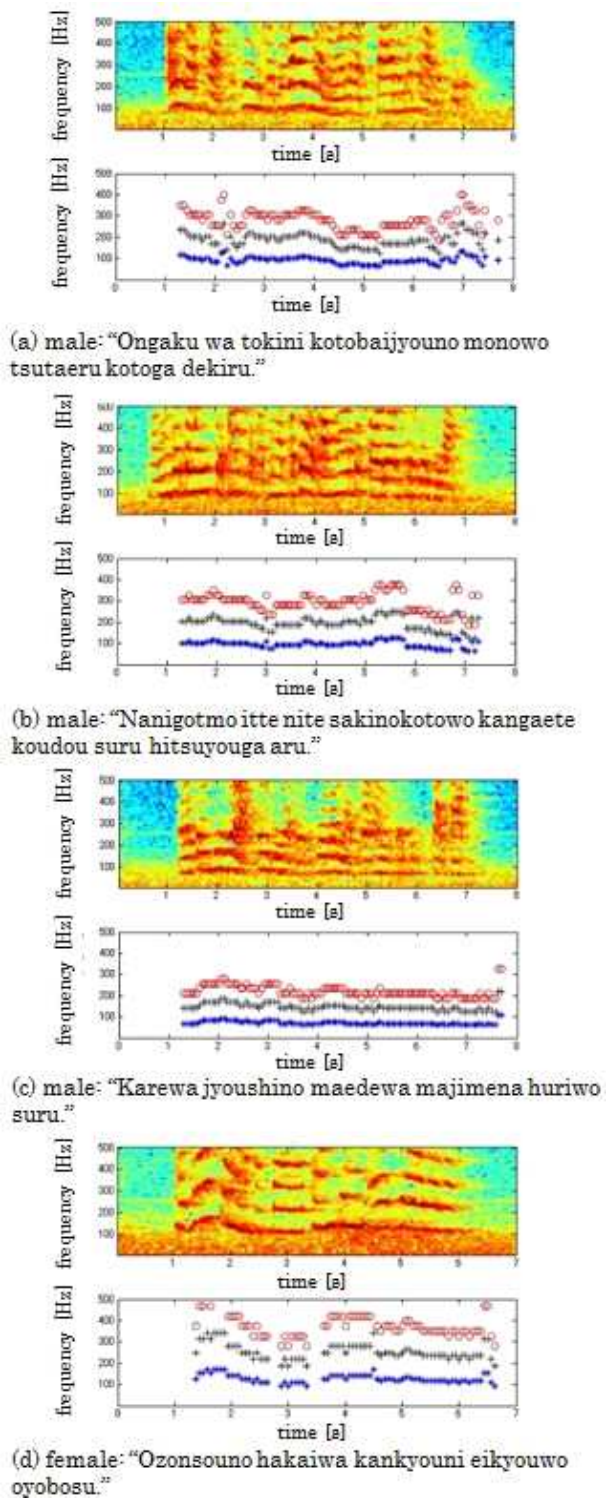


Fig. 7. GER and FER of the conventional and the proposed methods

4.2. Experiments under environmental noise. In order to confirm robustness of the proposed method against real environment noise, four speech signals in an office room are acquired (Fig. 8). Low frequency noises, such as air conditioner noise, exist in real room environment. That makes it difficult to estimate accurate $F0$, because $F0$ also exists in low frequency band.



(a) male: "Ongaku wa tokini kotobaijyouno monowo tsutaeru kotoga dekiru."

(b) male: "Nanigotmo itte nite sakinokotowo kangaete koudou suru hitsuyouga aru."

(c) male: "Karewa jyoushino maedewa majimena huriwo suru."

(d) female: "Ozonsouno hakaiwa kankyouni eikyouwo oyobosu."

Fig. 8. Estimation results of the proposed method in real environment noise (in each case: top chart is the spectrogram, bottom chart – results of $F0$, 2nd, 3d harmonics estimation)

In addition to low frequency noise, other factors are considered to cause failure estimation. One of these appears in Fig. 8b case. In this case, estimated spectrogram of speech does not always present exact harmonic (overtone) relationship. Figure 8b upper spectrogram at time interval around 4 [s] shows this phenomenon. Nevertheless, we can see at lower part of Fig. 8b that the proposed method can estimate almost accurate $F0$. For female speech result as shown in Fig. 8d, the estimated $F0$ is relatively accurate. This is because female $F0$ is usually higher than male $F0$, therefore, it is less degraded by lower frequency noise.

4.3. Speech feature evaluation. In the following experiments we evaluate the $F0$-based spectrogram normalization method, proposed in Sec. 2. The set of speech samples contained 200 spoken words and word sequences, coming from 4 speakers (2 male and 2 female speakers) [35]. The goal of experiments was to evaluate the similarity (we could also say: stability) of MFCC feature sets for every individual phoneme category, before and after the $F0$-based spectrogram normalization.

Voiced/unvoiced discrimination. Non-voiced parts are detected, when low-pass ratio is approximately below 0.4 and normalized autocorrelation is below 0.5. Silence is detected according to a power rule. Voiced parts are given when approximately the low-pass ratio is over 0.4 and the normalized auto-correlation is over 0.8 (see Figs. 9–11).

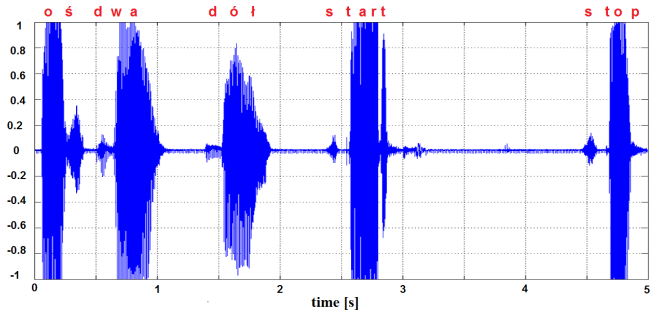


Fig. 9. The polish utterance: "*oś dwa dół start stop*"

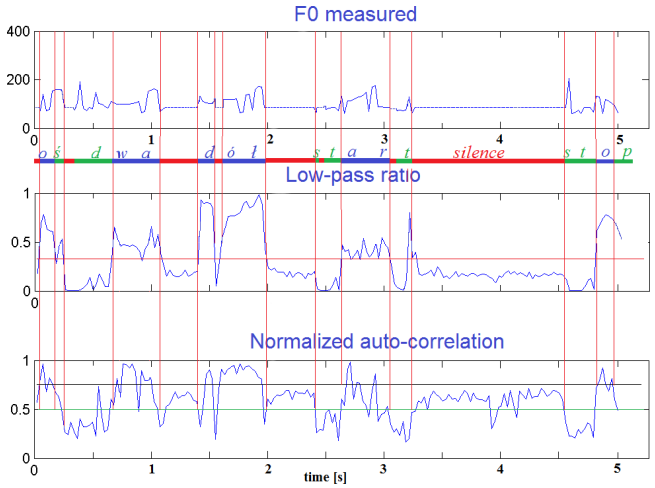


Fig. 10. Illustration of the measurement of F_0 in time (top drawing), the corresponding low-pass ratio distribution (middle drawing) and the normalized auto-correlation distribution (bottom drawing). The vertical lines illustrate the detected borders between voiced and unvoiced phonemes in the signal – they also provide time synchronization of all the drawings)

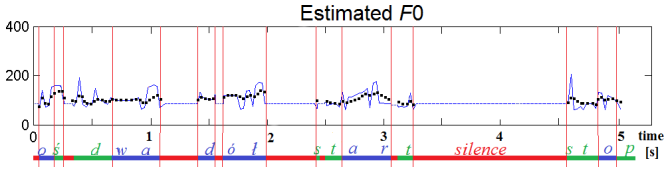


Fig. 11. Illustration of the estimated F_0. The dotted line represents the estimated F_0 distribution, whereas the continuous line represents the measurements of F_0 in time

Evaluation criteria. The MFCC features are computed according to the standard homeomorphism, i.e. the transformation [4]:

$$MFCC(x) = DFT^{-1}(\log(MFC(|DFT(x \cdot w)|))),$$

where x is the signal frame, w is the Hamming window, *MFC* is the transformation by a set of triangle band-pass filters located according to the Mel scale of frequencies, *DFT* and DFT^{-1} are the Discrete Fourier Transform and the Inverse DFT.

We managed manually to label them with the appearance of 12 selected phonemes: the vowels /a/, /e/, /o/; the approximants /y/, /r/, the nasal /n/, the fricatives /z/,/v/; the affricates /dZ/, /tS/; and the plosives /t/,/d/.

The stability of features for a single phoneme is expressed by two error measures. They represent the average square distance between two sets of feature vectors. Let $\mathbf{c}_N$ be a set of N feature vectors with L features each. $\mathbf{c}_N$ is compared with itself or with another set $\mathbf{C}_M$, that contains M vectors. The average within phoneme distance is:

$$\varepsilon_1(\mathbf{c}_N) = \frac{2}{N(N-1)} \sum_{m=1}^{N-1} \sum_{n=m+1}^{N} \frac{1}{L} \sum_{l=1}^{L} (c_l^n - c_l^m)^2. \quad (22)$$

The average distance between two sets $\mathbf{c}_N$ and $\mathbf{C}_M$ is:

$$\varepsilon_2(\mathbf{c}_N, C_M) = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{1}{L} \sum_{l=1}^{L} (c_l^n - C_l^m)^2. \quad (23)$$

Similarity of features before correction. Now, let us observe the behavior of MFCC features for a given speaker and in average (Fig. 12, Table 1). For a vowel and a nasal strong differences between speakers exist, that seem to correspond to the differences of speaker's basic frequencies. Contrary, affricates and consonants show good between-speaker similarities. The analysis of results given in Table 1 leads to following conclusions:

- the best similarity (small within-phoneme distances and variances, and large between-phoneme distances) is shown by plosives and unvoiced affricates;
- the between-phoneme distances are sufficiently large if compared to within-phoneme distances for every single speaker.

This observation verifies that a normalization procedure should focus on voiced phonemes.

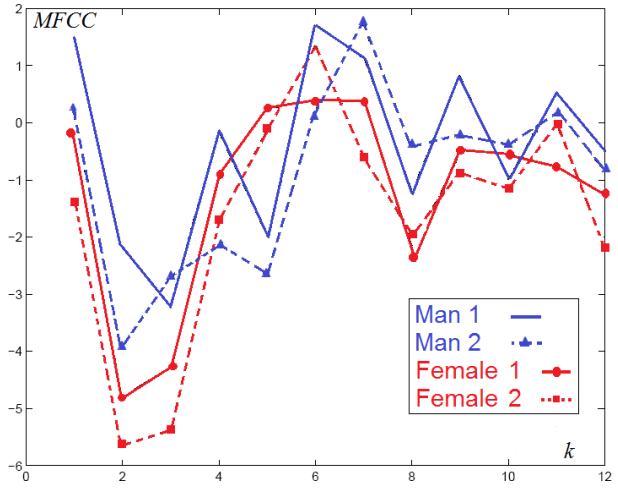


Fig. 12. Average MFCC feature vectors for vowel /a/ for 4 speakers

Table 1
Comparison of distances between features of phonemes for a single speaker and in average for all speakers

Speaker	Within-phoneme distance ε_1	Average distance ε_2 to other phonemes
Vowel /a/		
Male 1	2.22	19.2
Male 2	2.39	16.5
Female 1	2.42	19.5
Female 2	2.34	21.2
Average	2.28	19.1
Average for all speakers		
Vowel /e/	1.55	15.1
Vowel /o/	2.62	20.1
Approximant /y/	3.35	18.5
Approximant /r/	5.56	19.0
Nasal /n/	3.85	21.8
Fricative /z/	7.21	29.4
Fricative /v/	3.93	14.7
Affricate /dZ/	8.24	23.1
Affricate /tS/	3.20	40.5
Plosive /t/	1.93	14.3
Plosive /d/	1.82	13.8

Similarity of features after correction. We evaluate the average fundamental frequency for male utterances of studied phonemes to be around 12 Hz, whereas for female utterances this average value is around 200 Hz. In our normalization experiments the male and female utterances have been normalized separately. Firstly, male voices have been normalized to female average, hence $f_{F0-\text{ norm}}$ frequency was set to 200 Hz. Secondly, female voices have been normalized to the average of men, i.e. 12 Hz. The results of such two experiment series are summarized in Table 2. The relative change of average distance of all samples for given phoneme, before and after the normalization step, is obtained as:

$$\Delta\varepsilon_1(C_{N+M}) = \frac{\varepsilon_1^{(before)} - \varepsilon_1^{(after)}}{\varepsilon_1^{(before)}} \cdot 100\,\%. \quad (24)$$

Table 2

Total results of correcting female utterances by normalization to $f_{F0-\text{norm}} = 120$ Hz, and of correcting male utterances by normalization to $f_{F0-\text{norm}} = 200$ Hz

Phoneme	After female correction average $\Delta\varepsilon_1$:	After male correction average $\Delta\varepsilon_1$:
Vowels		
/a/	18.0 %	16.9 %
/e/	15.5 %	6.7 %
/o/	4.8 %	6.5 %
Approximants		
/y/	15.6 %	17.6%
/r/	0.9 %	−2.8%
Affricates		
/tS/	2.4 %	−0.1 %
/dZ/	4.5 %	12.5 %
Nasal		
/n/	−5.6 %	−1.0 %
Fricatives		
/v/	1.9 %	3.5 %
/z/	5.4 %	9.4 %
Plosives		
/d/	4.0 %	6.8 %
/t/	12.6 %	4.3 %
AVERAGE	5.2 %	6.7 %

Particular results of spectrogram normalization for /y/ (female voice) and /dZ/ (male voice) are illustrated in Figs. 13 and 14 Our approach allows to improve the stability of MFCC features generated for vowels (e.g. /a/, /e/, /o/) and voiced consonants (e.g. /y/). The articulation of these phonemes is not disturbed and attenuated by the vocal tract. In contrast, the retroflex /r/ is heavily attenuated and the normalization scheme offers no improvement for it. A positive influence onto the feature stability can also be observed for voiced fricatives and affricates (/z/ and /dZ/), and for plosives (/d/ and /t/). For nasal /n/, in contrary, there appears a small deterioration.

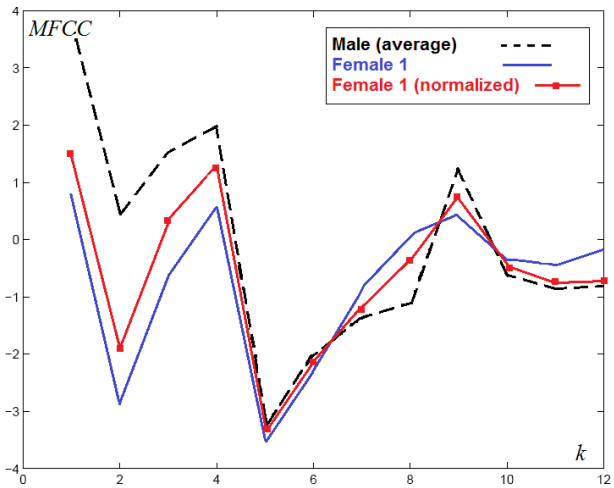


Fig. 13. Normalizing female speech features (MFCC) for approximant /y/ by mapping its fundamental frequency to default man's $F0$

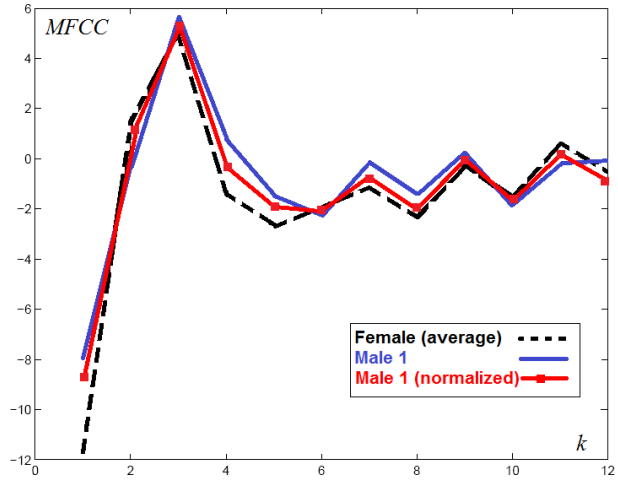


Fig. 14. Normalizing male speech features (MFCC) for affricate /dZ/ by mapping its fundamental frequency to default female's $F0$

5. Conclusions

The paper proposed a non-parametric recursive estimation of $F0$ and 2nd and 3rd harmonic frequencies utilizing particle filter. The $F0$ and its lower harmonics are estimated by using different likelihood function from the conventional. Our method has been compared with other conventional methods and has been proved to be more robust against background noise.

The obtained $F0$ and the 2nd and 3rd frequencies may not only be useful for speech recognition but can also be additional cues in speaker separation and localization [36, 37], many-channel speech deconvolution [38] and other speech processing tasks. It is because the direction or delay estimation is not always accurate especially in lower frequency band containing $F0$ and their harmonic frequencies.

To justify the approach, we have proposed an on-line spectrogram normalization scheme dedicated to improve the speaker-independency of the standard MFCC speech features. The approach relies of the highly reliable approach to the speaker's instantaneous fundamental frequency ($F0$) estimation. The advantage of our simple spectrogram normalization method is that it is computed "on-line" as there is no need for collecting particular speaker's samples in advance.

The original MFCC features are relatively stable for unvoiced and heavily attenuated phonemes, hence for such phonemes a normalization scheme is not necessary. For open or voiced phonemes the proposed approach was demonstrated to decrease the within phoneme average distance by 5–6%

REFERENCES

[1] J. Benesty, M.M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, Berlin, 2008.

[2] G. Demenko, B. Möbius, and K. Klessa, "Implementation of Polish speech synthesis for the BOSS system", *Bull. Pol. Ac.: Tech.* 58 (3), 371–376 (2010).

[3] M.M. Goodwin, "The STFT, sinusoidal models, and speech modification". in: *Springer Handbook of Speech Processing*, pp. 229–258, Springer, Berlin, 2008.

[4] U. Glavitsch: "Speaker normalization with respect to $F0$: a perceptual approach", in: *TIK-Report No. 185*, Eidgenössische Technische Hochschule Zürich, Zürich, 2003.

[5] D.O'Shaughnessy, "Formant estimation and tracking", in: *Springer Handbook of Speech Processing*, pp. 213–227, Springer, Berlin, 2008.

[6] R.W. Schafer, "Homomorphic systems and cepstrum analysis of speech", in: *Springer Handbook of Speech Processing*, pp. 161–180, Springer, Berlin, 2008.

[7] W.J. Hess, "Pitch and voicing determination", in: *Advances in Speech Signal Processing*, eds. S. Furui and M.M. Sondhi, pp. 3–48, Marcel Dekker. Inc., New York, 1992.

[8] A. de Cheveign'e and H. Kawahara, "Comparative evaluation of F0 estimation algorithms", *Proc. Eurospeech* 1, 2451–2454 (2001).

[9] M. Unoki and T. Hosorogiya, "Estimation of fundamental frequency of reverberant speech by utilizing complex cepstrum analysis", *J. Signal Processing* 12 (1), 31–44 (2008).

[10] H. Kawahara, H. Katayose, A. de Cheveign'e and R.D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity", *Proc. Eurospeech* 1999, 2781–2784 (1999).

[11] A. de Cheveign'e and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music", *J. Acoust. Soc. Am.* 111 (4), 1917–1930 (2002).

[12] T. Miwa, Y. Tadokoro, and T. Saito, "The pitch estimation of different musical instruments sounds using comb filters for transcription", *IEICE Trans.* D-2, J81-D-2 (9), 1965–1974 (1998).

[13] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components", *J. Acoust. Soc. Am.*, 116 (6), 3690–3700 (2004).

[14] Y. Ishimoto, M. Unoki, and M. Akagi, "A fundamental frequency estimation method for noisy speech based on instantaneous amplitude and frequency", *Proc. EuroSpeech* 2001, 2439–2442 (2001).

[15] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, and K. Shinkano, "Robust estimation of fundamental frequency using instantaneous frequencies of harmonic components", *IEICE Proc.* D-2, J83-D-2 (11), 2077–2086 (2000).

[16] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: a flexible bayesian approach", *IEEE Trans. on Audio Speech and Language Processing* 15 (4), 1283–1295 (2007).

[17] S. Kim, A.S. Paul, E.A. Wan, and J. McNames, "Multiharmonic tracking using sigmapoint Kalman filter", *IEEE EMBC* 8, CD-ROM (2008).

[18] K. Nishi, M. Abe, and S. Ando, "Multiple pitch tracking and harmonic segregation algorithm for auditory scene analysis", *The Society of Instrument and Control Engineers* 34 (6), 483–490 (1988), (in Japanese).

[19] S. Hainsworth and M. Macleod, "Beat tracking with particle filtering algorithms", *Proc. WASPAA* 1, 91–94 (2003).

[20] S. Tomoike and M. Akagi, "Estimation of local peaks based on particle filter in advance environments", *J. Signal Processing* 12 (4), 303–306 (2008).

[21] L. Lee and R. Rose, "A frequency warping approach to speaker normalization", *IEEE Trans. on Speech and Audio Processing* 6 (1), 49–60 (1998).

[22] P. Dognin, "A bandpass transform for speaker normalization", *Ph.D. Dissertation*, University of Pittsburgh, Pittsburgh, 2003.

[23] H. Traunmüller and F. Lacerda, "Perceptual relativity in identification of two-formant vowels", *Speech Communication* 6, 143–157 (1987).

[24] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization", *Proc. ICASSP* 1, 346–348 (1996).

[25] J. Laroche and M. Dolson, "New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing, and other exotic audio modifications", *J. Audio Eng. Soc*. 47 (11), 928–936 (1999).

[26] L.R. Rabiner, "On the use of autocorrelation analysis for pitch", *IEEE Trans. on Acoustics, Speech, and Signal Processing* ASSP-25 (1), 24–33 (1997).

[27] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech", *IEEE Trans. on Speech and Audio Processing* 9 (7), 727–730 (2001).

[28] G.S. Ying, L.H. Jamieson, and C.D. Mitchell, "A probabilistic approach to AMDF pitch detection", *J. Acoust. Soc. Am.* 95 (5), 2817–2817 (1994).

[29] T. Miyamoto, H. Inada, and K. Nakata, "A real time PARCOR analysis of speech by high- performance signal processors", *IEICE* J66-A (7), 625–632 (1983), (in Japanese).

[30] T. Sakai, T. Kitamura, and E. Hayahara, "Improvement of pitch extraction method in noisy environment based on cepstrum", *Electronics, Information, and Communication Engineers* 1, 299 (1995).

[31] D.M. Haward, "Peak-picking fundamental period estimation for hearing prostheses", *J. Acoust. Soc. Am*. 86 (3), 902–910 (1989).

[32] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter. Particle Filters for Tracking*, Artech House, DSTO, Boston, 2004.

[33] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech", *IEEE Trans. on Signal Processing* Vil. 39 (1), CD-ROM (1991).

[34] P. Veprek and M.S. Scordilis, "Analysis, enhancement and evaluation of five pitch determination techniques", *Speech Comm*. 37, 249–270 (2002).

[35] B. Adamczyk, K. Adamczyk, and K. Trawiński, "Robot's vocabluary", *IAiR Bulletin* 12, CD-ROM (2000), (in Polish).

[36] G.-N. Hu and D.-L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation", *IEEE Trans. on Neural Networks* 15 (5), 1135–1150 (2004).

[37] W. Kasprzak, N. Ding, and N. Hamada: "Relaxing the WDO assumption in blind extraction of speakers from speech mixtures", *J. Telecom. and Information Technology* 4, 50–58 (2010).

[38] F.A. Okazaki and W. Kasprzak: "A two-step approach to blind deconvolution of speech and sound sources in the time domain", *Bull. Pol. Ac.: Tech.* 53 (1), 49–55 (2005).