Limited error based event localizing temporal decomposition and its application to variable-rate speech coding

Phu Chien Nguyen [*], Masato Akagi, Binh Phu Nguyen

Graduate School of Information Science,
Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292, JAPAN

Abstract

This paper proposes a novel algorithm for temporal decomposition (TD) of speech, called 'Limited Error Based Event Localizing Temporal Decomposition' (LEBEL-TD), and its application to variable-rate speech coding. In previous work with TD, TD analysis was usually performed on each speech segment of about 200-300 ms or more, making it impractical for online applications. In this present work, the event localization is determined based on a limited error criterion and a local optimization strategy, which results in an average algorithmic delay of 65 ms. Simulation results show that an average log spectral distortion of about 1.5 dB can be achievable at an event rate of 20 events/sec. Also, LEBEL-TD uses neither the computationally costly singular value decomposition routine nor the event refinement process, thus reducing significantly the computational cost of TD. Further, a method for variable-rate speech coding an average rate of around 1.8 kbps based on STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum), which is a high-quality speech analysis-synthesis framework, using LEBEL-TD is also realized. Subjective test results indicate that the performance of the proposed speech coding method is comparable to that of the 4.8 kbps FS-1016 CELP coder.

Key words: Temporal decomposition; Event vector; Event function; STRAIGHT; Speech coding; Line spectral frequency.

[*] Corresponding author, currently with the Institute of Scientific and Industrial Research, Osaka University, Japan.

Email addresses: `chien@ar.sanken.osaka-u.ac.jp` (Phu Chien Nguyen), `akagi@jaist.ac.jp` (Masato Akagi), `npbinh@jaist.ac.jp` (Binh Phu Nguyen).

1 Introduction

Most existing low rate speech coders analyze speech in frames according to a model of speech production. Such a model is the linear predictive coding (LPC) model. However, speech production can be considered as a sequence of overlapping articulatory gestures, each producing an acoustic event that should approximate an articulatory target (Fallside and Woods, 1985). Due to co-articulation and reduction in fluent speech, a target may not be reached before articulation towards the next phonetic target begins. The non-uniform distribution of these speech events is not exploited in frame-based systems.

The so-called temporal decomposition (TD) method (Atal, 1983) for analyzing the speech signals achieves the objective of decomposing speech into targets and their temporal evolutionary patterns without any recourse to any explicit phonetic knowledge. This model of speech takes into account the above articulatory considerations and results in a description of speech in terms of a sequence of overlapping event functions and corresponding event vectors as given in Equation (1).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \qquad 1 \leq n \leq N \tag{1}$$

where $\mathbf{a}_k$, the k^{th} event vector, is the speech parameters corresponding to the k^{th} target. The temporal evolution of this target is described by the k^{th} event function, $\phi_k(n)$. $\hat{\mathbf{y}}(n)$ is the approximation of $\mathbf{y}$(n), the n^{th} spectral parameter vector, produced by the TD model. N and K are the number of frames and number of events in the block of spectral parameters under consideration, respectively.

Despite the fact that TD has the potential to become a versatile tool in speech analysis, its high-computational complexity and long-algorithmic delay make it impractical for online applications. In the original TD method by Atal (1983), TD analysis is performed on each speech segment of about 200-300 ms, thus resulting in an algorithmic delay of more than 200 ms. In addition, Atal's method is very computationally costly, which has been mainly attributed to the use of the singular value decomposition (SVD) routine and the iterative refinement process (Van Dijk-Kappers and Marcus, 1989). These prevent Atal's method from online applications.

Most of modified algorithms for TD have been mainly proposed to overcome the drawback of high computational cost incurred by the original TD method. The algorithm for TD proposed in (Nandasena et al., 2001), S^{2}BEL-TD, reduces the computational cost of TD by avoiding the use of SVD, but the long algorithmic delay has more or less remained the same. S^{2}BEL-TD uses a spectral stability criterion to determine the initial event locations. Meanwhile, the event localization in the optimized TD (OTD) method (Athaudage et al., 1999) is performed using an optimized approach (dynamic programming). Although the OTD method can achieve very good results in terms of reconstruction accuracy, but its long algorithmic delay (more than 450 ms) makes it suitable for speech storage related applications only. Also, both the OTD and S^{2}BEL-TD methods use the line spectral frequency (LSF) parameters (Itakura, 1975) as input, which might cause the corresponding LPC synthesis filter to be unstable. This is because there is no guarantee that the selected LSF parameters are valid after the spectral transformation performed by these two TD methods. The restricted TD (RTD) (Kim and Oh, 1999) and the modified RTD (MRTD) (Nguyen and Akagi, 2002a) methods, on the other hand, consider the ordering property of LSFs to make LSF parameters possible for TD. These methods require an average algorithmic delay of about 95 ms, while can achieve relatively good results.

In this paper, we propose a novel algorithm for temporal decomposition of speech called 'Limited Error Based Event Localizing Temporal Decomposition' (LEBEL-TD). This method employs the restricted second order model and a novel approach to event localization. Here, the event localization is initially performed based on a limited error criterion, and then further refined by a local optimization strategy. In the following, the event vectors are set as the original spectral parameter vectors at the event locations and thus, it can be applied to decomposing the LSF parameters without considering the ordering property of LSFs. This algorithm for TD requires only 65 ms average algorithmic delay [1], while can achieve results comparable to the S^{2}BEL-TD, RTD and MRTD methods. Moreover, LEBEL-TD uses neither the computationally costly SVD routine nor the iterative refinement process, thus resulting in a very low computational cost required for TD analysis.

We have also investigated the usefulness of LEBEL-TD in speech coding. In this paper, a method for variable-rate speech coding based on STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) (Kawahara et al., 1999), which is a high-quality speech analysis-synthesis framework, using LEBEL-TD is presented. For encoding spectral information of speech, LEBEL-TD based vector quantization (VQ) is utilized, whilst other speech parameters are quantized using scalar quantization (SQ). Subjective results indicate that the performance of the proposed speech coding method at an average rate of around 1.8 kbps can be comparable to that of the 4.8 kbps US Federal Standard FS-1016 CELP coder (Campbell et al., 1991).

[1] the frame period is 10 ms long, as considered in S^{2}BEL-TD, MRTD, and OTD.

2 LEBEL-TD of speech spectral parameters

2.1 Restricted second order TD model

Assume that the co-articulation in speech production described by the TD model in terms of overlapping event functions is limited to adjacent events, the second order TD model (Niranjan and Fallside, 1989; Shiraki and Honda, 1991; Athaudage et al., 1999), where only two adjacent event functions can overlap as depicted in Fig. 1, is given by Equation (2).

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k\phi_k(n) + \mathbf{a}_{k+1}\phi_{k+1}(n), \;\; n_k \le n < n_{k+1} \tag{2}$$

where n_k and n_{k+1} are the locations of event k and event $(k+1)$, respectively.

(Fig. 1 is around here)

The so-called restricted second order TD model was utilized in (Dix and Bloothooft, 1994; Kim and Oh, 1999; Nguyen and Akagi, 2002a,b) and this work with an additional restriction to the event functions in the second order TD model that all event functions at any time sum up to one. The argument for imposing this constraint on the event functions can be found in (Dix and Bloothooft, 1994). Equation (2) can be rewritten as follows.

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k\phi_k(n) + \mathbf{a}_{k+1}\left(1 - \phi_k(n)\right), \;\; n_k \le n < n_{k+1} \tag{3}$$

2.2 Determination of event functions

Assume that the locations n_k and n_{k+1} of two consecutive events are known. Then, the right half of the k^{th} event function and the left half of the $(k+1)^{th}$ event function can be optimally evaluated by using $\mathbf{a}_k = \mathbf{y}(n_k)$ and $\mathbf{a}_{k+1} = \mathbf{y}(n_{k+1})$. The reconstruction error, $E(n)$, for the n^{th} spectral parameter vector is

$$\begin{aligned} E(n) &= \parallel \mathbf{y}(n) - \hat{\mathbf{y}}(n) \parallel^2 \\ &= \parallel (\mathbf{y}(n) - \mathbf{a}_{k+1}) - (\mathbf{a}_k - \mathbf{a}_{k+1})\phi_k(n) \parallel^2 \end{aligned} \tag{4}$$

where, $n_k \le n < n_{k+1}$. Therefore, $\phi_k(n)$ should be determined so that $E(n)$ is minimized.

2.2.1 Geometric interpretation of TD

TD yields an approximation of a sequence of spectral parameters by a linear combination of event vectors. Since TD's underlying distance metric is Euclidean, a natural requirement is to have this approximation be invariant with respect to a translation or rotation of the spectral parameters. Dix and Bloothooft (1994) considered the geometric interpretation of TD results and found that TD is rotation and scale invariant, but it is not translation invariant.

In order to overcome this shortcoming and describe TD as a breakpoint analysis procedure in a multidimensional vector space, where breakpoints are connected by straight line segments, Dix and Bloothooft (1994) enforced two constraints on the event functions: (i) at any moment of time only two event functions, which are adjacent in time, are non-zero; and (ii) all event functions at any time sum up to one. In other words, the restricted second order TD model was utilized in (Dix and Bloothooft, 1994). These constraints are needed to approximate the path in parameter space by means of straight line segments between breakpoints (see Fig. 2).

(Fig. 2 is around here)

Geometrically speaking, the two event vectors $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ define a plane in P-dimensional vector space. The determination of event functions $\phi_k(n)$ and $\phi_{k+1}(n)$ in the interval $[n_k, n_{k+1}]$ is now depicted in Fig. 3(a) as the projection of vector $y(n)$ onto this plane. Clearly the following holds: $\phi_k(n_k) = 1$, $\phi_k(n_{k+1}) = 0$, and $0 \leq \phi_k(n) \leq 1$ for $n_k \leq n \leq n_{k+1}$.

(Fig. 3 is around here)

While n ranges from n_k to n_{k+1}, the movement of vector $\mathbf{y}(\mathbf{n})$ is described by the transition of $\hat{\mathbf{y}}(n)$ along the straight line segment connecting two breakpoints $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$. As time is moving forward, the transition of $\hat{\mathbf{y}}(n)$ should be monotonic.

2.2.2 New determination of event functions

The TD model is based on the hypothesis of articulatory movements towards and away from targets. An appealing result of the above properties of event functions is that one can interpret the values $\phi_k(n)$ as a kind of activation values of the corresponding event. During the transition from one event towards the next the activation value of the left event decreases from one to zero, whilst the right event increases its activation value from zero to the value of one. As mentioned earlier, to model the temporal structure of speech more effectively no backwards transitions are allowed. Therefore, each event function should have a growth cycle; during which the event function grows from zero to one and a decay cycle; during which the event function decays from one to zero. In other words, each event function should have only one peak, which is called the well-shapedness property. On the contrary, an ill-shaped event function can be viewed as an event function which has several growth and decay cycles, i.e. having more than one peak.

Fig. 4 shows examples of well-shaped and ill-shaped event functions. It can be seen that well-shaped event functions are desirable from speech coding point of view because the well-shapedness property helps reduce the quantization error of event functions when vector quantized.

(Fig. 4 is around here)

However, the determination of event functions in (Dix and Bloothooft, 1994) has not guaranteed the well-shapedness property for them since their changes during the transition from one event towards the next may not be monotonic, which results in ill-shaped event functions. In particular, one may wonder that if an event function has some values of one interlaced by other values, causing the next event function to have more than one lobe, which is not acceptable in the conventional TD method. Ill-shaped event functions are also undesirable from speech coding point of view. They increase the quantization error when vector quantized because the uncharacteristic valleys and secondary peaks are not normally captured by the codebook functions. This is because an event function is quantized by its length and shape in the interval between its and the next event function's locations. In that interval, a well-shaped event function is always a decreasing function while an ill-shaped event function is always non-monotonic.

Taking into account the above considerations, we have modified the determination of event functions corresponding to the point of the line segment between $\hat{\mathbf{y}}(n-1)$ and $\mathbf{a}_{k+1}$ (see Fig. 3(b)) instead of $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ as considered in (Dix and Bloothooft, 1994), with minimum distance from $\mathbf{y}(n)$. In mathematical form, the above determination of event functions can be written as

$$\phi_k(n) = \begin{cases} 1-\phi_{k-1}(n), & \text{if } n_{k-1} < n < n_k \\ 1, & \text{if } n = n_k \\ \min(\phi_k(n-1), \max(0, \hat{\phi}_k(n))), & \text{if } n_k < n < n_{k+1} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where

$$\hat{\phi}_k(n) = \frac{\langle (\mathbf{y}(n) - \mathbf{a}_{k+1}), (\mathbf{a}_k - \mathbf{a}_{k+1}) \rangle}{\| \mathbf{a}_k - \mathbf{a}_{k+1} \|^2} \tag{6}$$

Here, $< .,. >$ and $\| \, . \, \|$ denote the inner product of two vectors and the norm of a vector, respectively.

This modification ensures that the value of event function ϕ_k at n is always not greater than the value of event function ϕ_k at $n-1$ in the interval $[n_k; n_{k+1}]$ (see the third line of Equation (5)), and thereby the well-shapedness property is guaranteed. It should be noted that in (Dix and Bloothooft, 1994), $\phi_k(n)$ is determined as $\min(1, \max(0, \hat{\phi}_k(n))$, if $n_k < n < n_{k+1}$.

2.3 LEBEL-TD algorithm

The section of spectral parameters, $\mathbf{y}(n)$, where $n_k \leq n < n_{k+1}$, is termed a segment. The total accumulated error, $E_{seg}(n_k, n_{k+1})$, for the segment is

$$E_{seg}(n_k, n_{k+1}) = \sum_{n=n_k}^{n_{k+1}-1} E(n) \qquad (7)$$

where, $E(n)$ can be calculated for every $n_k \leq n < n_{k+1}$ using Equation (4) once n_k and n_{k+1} are known. The buffering technique for LEBEL-TD is depicted in Fig. 5, and the whole algorithm is described as follows.

Step 0. Set $k \leftarrow 1$, $n_1 \leftarrow 1$, $a_1 \leftarrow \mathbf{y}(1)$; set n_2 as the last location from n_1 on so that the reconstruction error for every frame in the interval (n_1, n_2) is less than a predetermined number ε.

Step 1. Similarly, set n_3 as the last location from n_2 on so that the reconstruction error for every frame in the interval (n_2, n_3) is less than ε.

Step 2. Local optimize the location of n_2 in the interval (n_1, n_3).

$$n_2^* = \arg \min_{n_1 < n_2 < n_3} \{E_{seg}(n_1, n_2) + E_{seg}(n_2, n_3)\}$$

where, only n_2 that makes $E(n) < \varepsilon$ for every $n_1 < n < n_3$ is taken into account. If n_3 is the last frame, set $k \leftarrow k+1$, $a_k \leftarrow \mathbf{y}(n_2^*)$, $a_{k+1} \leftarrow \mathbf{y}(n_3)$; and exit.

Step 3. Set $k \leftarrow k+1$, $a_k \leftarrow \mathbf{y}(n_2^*)$; then set $n_1 \leftarrow n_2^*$, $n_2 \leftarrow n_3$; and go back to step 1.

(Fig. 5 is around here)

The predetermined number ε is called the reconstruction error threshold, and it is the only parameter that effects the number and locations of the events. The reconstruction error threshold controls the event rate, i.e. the number of events per second, and can be appropriately selected to achieve the optimal performance of TD analysis for different applications. This is the reason why the above TD algorithm is named 'Limited Error Based Event Localizing Temporal Decomposition' (LEBEL-TD).

$\varepsilon = 0.045$ was empirically chosen as a suitable value for the reconstruction error threshold to produce the event rate of about 20 events/sec. It should be noted that the spectral parameter here is LSF with the frame period is set as 10 ms. This event rate results in an average buffering delay of about 50 ms, i.e. 5 frames, along with a 10 ms, i.e. one frame, look-ahead. On the other hand, the LPC analysis window is 30 ms long, which implies a 15 ms look-ahead. The calculation of algorithmic delay for LEBEL-TD is depicted in Fig. 6. Note that this calculation is applied to analyzing the first segment. From the second segment on, the look-ahead frame in the last segment can be employed. Therefore, the average algorithmic delay for LEBEL-TD is about 65 ms and has been known to be the lowest algorithmic delay for TD so far. Moreover, LEBEL-TD has significantly reduced the computational cost of TD because it uses neither the computationally costly SVD routine nor the iterative refinement process. These make LEBEL-TD suitable for online applications.

(Fig. 6 is around here)

In the LEBEL-TD method, the event vectors are set as the spectral parameter vectors corresponding to the event locations. Obviously, the event vectors are valid spectral parameter vectors and the stability of the corresponding LPC synthesis filter can be thus ensured after spectral transformation performed by LEBEL-TD. Consequently, LEBEL-TD can be applied to analyzing any current types of parametric representations of speech. Meanwhile, most conventional TD methods use an iterative refinement of event vectors, which might cause the reconstructed spectral parameter vectors to be invalid, for example when TD is applied to decomposing LSF parameters, and thus resulting in an unstable LPC synthesis filter.

Fig. 7 shows the plot of event functions obtained from LEBEL-TD analysis of LSF parameters for an example of a female/Japanese sentence utterance 'shimekiri ha geNshu desu ka.' As can be seen from the figure, all event functions are well-shaped. In Fig. 8, the plots of original and reconstructed LSF parameters after LEBEL-TD analysis are shown for the same utterance as utilized in Fig. 7.

(Fig. 7 is around here)

(Fig. 8 is around here)

3 Performance evaluation

The ATR Japanese speech database was used for the speech data. Line spectral frequency (LSF) parameters introduced by Itakura (1975) have been selected as the spectral parameter for the LEBEL-TD. This is because it is well-known that LSF parameters have the best interpolation (Paliwal, 1995) and quantization (Paliwal and Atal, 1993) properties over the other LPC related spectral parameters.

Log spectral distortion (LSD) is a commonly used measure in evaluating the performance of LPC quantization (Paliwal and Atal, 1993) and interpolation (Paliwal, 1995). LSD measure is also used for evaluating the interpolation performance of TD algorithms (Shiraki and Honda, 1991; Athaudage et al., 1999; Nandasena et al., 2001). This criterion is a function of the distortion introduced in the spectral density of speech in each particular frame. Log spectral distortion, D_n, for the nth frame is defined (in dB) as follows.

$$D_n = \sqrt{\frac{1}{F_s}\int\limits_0^{F_s}[10log_{10}(P_n(f)) - 10log_{10}(\hat{P}_n(f))]^2 df}$$

where F_s is the sampling frequency, and $P_n(f)$ and $\hat{P}_n(f)$ are the LPC power spectra corresponding to the nth frame of the original spectral parameters, $\mathbf{y}(n)$, and the reconstructed spectral parameters, $\hat{\mathbf{y}}(n)$, respectively. The results are provided in terms of log spectral distortion histograms, average log spectral distortion and percentage outliers having log spectral distortion greater than 2 dB. The outliers are divided into the following two types. Type 1: consists of outliers in the range 2-4 dB, and Type 2: consists of outliers having spectral distortion greater than 4 dB.

A set of 250 sentence utterances of the ATR Japanese speech database were selected as the speech data. This speech dataset consists of about 20 minutes of speech spoken by 10 speakers (5 male & 5 female) re-sampled at 8 kHz sampling frequency. 10^{th} order LSF parameters were calculated using a LPC analysis window of 30 ms at 10 ms frame intervals, and LEBEL-TD analyzed. Additionally, log spectral distortion was also evaluated over the same speech dataset for three other methods of TD: S^{2}BEL-TD (Nandasena et al., 2001), RTD (Kim and Oh, 1999), and MRTD (Nguyen and Akagi, 2002a) with LSF as the spectral parameter. The event rate was set as around 20 events/sec for all the four methods.

Table 1 gives a comparison of the log spectral distortion results for the LEBEL-TD, S^{2}BEL-TD, RTD, and MRTD algorithms. The distribution of the log spectral distortion in the form of histograms is shown in Fig. 9. Results indicate slightly better performance in the case of S^{2}BEL-TD over the others, followed by LEBEL-TD and then RTD. However, it has been shown in (Nguyen and Akagi, 2002a) that the S^{2}BEL-TD and RTD methods, in the current forms, cannot always be applied to analyzing LSF parameters due to the stability problems in the LPC synthesis filter. Also, LEBEL-TD requires a lower computational cost for TD analysis than MRTD, which is mainly attributed to the fact that LEBEL-TD does not employ the iterative refinement process. In addition, LEBEL-TD also needs a shorter algorithmic delay than MRTD. For these reasons, LEBEL-TD hereafter is used for analyzing the LSF parameters.

(Table 1 and Fig. 9 are around here)

We have also evaluated the performance of LEBEL-TD on the above speech dataset for some ε. Table 2 gives the summary of LSD and the event rate obtained from LEBEL-TD analysis for some different values of ε. As can be seen from the table, the event rate decreases and the average LSD increases as ε increases. Fig. 10 illustrates the average log spectral distortion versus the event rate.

(Table 2 and Fig. 10 are around here)

It is clear that the event rate controls the delay. The event rate also controls the coding quality and the bit-rate. Fig. 11 shows an example of average log spectral distortion versus average algorithmic delay obtained from LEBEL-TD analysis of the above speech dataset. It is demonstrated that the longer the algorithmic delay, the larger the log spectral distortion.

(Fig. 11 is around here)

4 Variable-rate speech coding based on STRAIGHT using LEBEL-TD

As shown earlier, in the temporal decomposition (TD) framework, the speech is no longer represented by a vector updated frame by frame, but instead by the continuous trajectory of a vector. The trajectory is decomposed into a set of phoneme-like events, i.e. a series of temporally overlapping event functions and a corresponding series of event vectors. Since the event rate varies in time, TD can be considered as a technique to be used for variable-rate speech coding.

The linear predictive coding (LPC) model of speech has been widely adopted in many speech coding systems (Campbell and Tremain, 1991; Campbell et al., 1991; Paliwal and Atal, 1993; Paliwal, 1995). However, since the line spectral frequency (LSF) parameters derived from LPC analysis are independently extracted on a frame-by-frame basis, the corresponding LSF parameter vector trajectory is not so smooth. In the other case, STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum), invented by Kawahara et al. (1999), can extract very smooth spectrogram by employing a time-frequency interpolation procedure. As a consequence, LSF parameters extracted from the spectrogram are correlated among frames, and the corresponding LSF parameter vector trajectory is thus smooth, which is desirable for TD algorithms. In addition, STRAIGHT is known as a high-quality speech analysis-synthesis framework with many promising applications in speech synthesis and modification (Kawahara et al., 1999). This versatile speech manipulation toolkit can roughly decompose input speech signals into spectral envelopes, i.e. spectrogram, F0 (fundamental frequency) information, and noise ratios. Those parameters and the maximum value of amplitude are required for resynthesizing high-quality speech. To make STRAIGHT applicable for low-bit-rate speech coding, the bit rate required to represent the spectral envelope must be minimized. Since the spectral envelopes can be further analyzed into LSF parameters and gain information, the LEBEL-TD algorithm can be incorporated with STRAIGHT to construct high-quality speech coders working at low-bit rates.

In this section, we introduce a new method for variable-rate speech coding based on STRAIGHT using LEBEL-TD. The proposed speech encoder and decoder block diagrams are shown in Fig. 12, and a detailed description of the proposed speech coding method is shown in the subsections followed. Experimental results show that this speech coding method can produce good quality speech at an average rate below 2 kbps.

(Fig. 12 is around here)

4.1 Derivation of LSF Parameters

The amplitude spectrum $X[m]$, where $0 \leq m \leq \frac{M}{2}$ with M is the number of samples in the frequency domain, obtained from STRAIGHT analysis is transformed to the power spectrum using Equation (8).

$$S[m] = \mid X[m] \mid^2, \quad 0 \leq m \leq \frac{M}{2} \tag{8}$$

The i^{th} autocorrelation coefficient, R[i], is then calculated using the inverse Fourier transform of the power spectrum as follows.

$$R[i] = \frac{1}{M} \sum_{m=0}^{M-1} S[m] \exp\{j \frac{2\pi mi}{M}\}, \quad 0 \leq i \leq M-1 \tag{9}$$

where $S[m] = S[M-m]$. Assume that the speech samples can be estimated by a P-th order all-pole model, where $0 < P < M$, the reconstruction error is calculated as given in Equation (10).

$$P_L = R[0] - \sum_{l=1}^{P} a_l^P R[l] \tag{10}$$

where $\{a_l^P\}$, $l = 1, 2 \cdots P$, are the corresponding linear predictive coding (LPC) coefficients. P_L hereafter is referred to as gain. By minimizing P_L with respect to a_l^P, where $l = 1, 2 \cdots P$, a_l^Ps could be evaluated. They are then transformed to the LSF parameters.

4.2 LEBEL-TD based vector quantization of LSF Parameters

4.2.1 Vector quantization of event vectors

Since the event vectors obtained from LEBEL-TD method are valid LSF parameter vectors, they can be quantized by usual quantization methods for LSF parameters. Here, the split vector quantization introduced in (Paliwal and Atal, 1993) was adopted. In this work, the order of LSFs was empirically selected as 32 to increase the quality of reconstructed speech. Every event vector was divided into four subvectors of dimensions 7, 8, 8, 9 due to the distribution of LSFs, and each subvector was quantized independently. We assigned 8 bits to each subvector, which resulted in 32 bits allocated to one event vector.

4.2.2 Vector quantization of event functions

In the case of event functions, normalization of the event functions is necessary to fix the dimension of the event function vector space. Notice that only quantizing $\phi_k(n)$ in the interval $[n_k; n_{k+1}]$ is enough to reconstruct the whole event function $\phi_k(n)$. Moreover, $\phi_k(n)$ always starts from one and goes down to zero in that interval, and the type of decrease (after normalizing the length of $\phi_k(n)$) can be vector quantized. Therefore, an event function $\phi_k(n)$ can be quantized by its length $L(k) = n_{k+1} - n_k$ and shape in $[n_k + 1; n_{k+1} - 1]$.

In this work, 15 equidistant samples were taken from each event function for length-normalization and then vector quantized by a 7-bit codebook.

Considering that all intervals between two consecutive event locations are less than 256 frames long (note that the frame period used in STRAIGHT analysis is 1 ms long), we used 8 bits for quantizing the length of each event function. Shortly speaking, each $\phi_k(n)$ was quantized by its length and the type of decrease.

4.3 Coding speech excitation parameters

4.3.1 Coding noise ratio parameters

The speech production mechanism is assumed to be a synchronously controlled process with respect to the movement of different articulators, i.e. jaws, tongue, larynx, glottis etc., and the temporal evolutionary patterns of different properties of speech, e.g. spectrum, F0, and noise ratio, can be thus described by a common set of event functions (Nandasena et al., 2001). Therefore, the same event functions obtained from LEBEL-TD analysis of LSF parameters are also used to describe the temporal evolution of the noise ratio parameters. Let $i(n)$ be a noise ratio parameter. We have $0 \leq i(n) \leq 1$, where $i(n) = 1$ for white noise and $i(n) = 0$ for pure pulse. Then $i(n)$ is approximated by $\hat{i}(n)$, the reconstructed noise ratio parameter for the nth frame, as follows in terms of noise ratio targets, i_ks, and the event functions, $\phi_k(n)$s. Since the event functions are quantized and transmitted, this description also helps reduce the bit rate required for encoding noise ratio information.

$$\hat{i}(n) = \sum_{k=1}^{K} i_k \phi_k(n), \quad 1 \leq n \leq N$$

The noise ratio targets are determined by minimizing the sum squared error, E_i, between the original and the interpolated noise ratio parameters with respect to i_ks.

$$E_i = \sum_{n=1}^{N} \Big(i(n) - \hat{i}(n) \Big)^2 = \sum_{n=1}^{N} \Big(i(n) - \sum_{k=1}^{K} i_k \phi_k(n) \Big)^2$$

where, $i(n)$ is the original noise ratio parameter for the n^{th} frame. Finally, the noise ratio targets are quantized by using scalar quantization. In this work, we used 6 bits for quantizing each noise ratio target.

Fig. 13 shows the plots of original and reconstructed noise ratio parameters and the plot of frame-wise noise ratio error, $e_i(n)$, where $e_i(n) = \hat{i}(n) - i(n)$, for a male/Japanese sentence utterance. The root mean squared (RMS) noise ratio error, $\sqrt{E_i}$, where $E_i = \frac{1}{N}\sum_{n=1}^{N} {e_i}^2(n)$, was found to be about 0.1166.

(Fig. 13 is around here)

4.3.2 Coding F0 parameters

For encoding F0 information, the lengths of voiced and unvoiced segments were quantized by scalar quantization first, with an average bit rate of 36 bps. Next, linear interpolation was used within the unvoiced segments to form a continuous F0 contour. Similar to the method presented in subsection 4.3.1, the continuous F0 contour was then described by the event functions obtained from LEBEL-TD analysis of LSF parameters and the so-called F0 targets. As mentioned earlier, this description also helps reduce the bit rate required for encoding F0 information since we can make use of the encoded event functions. The F0 targets were then quantized by a 6-bit logarithmic quantizer. In the decoder, F0 values were reconstructed from the quantized event functions and F0 targets using the TD synthesis. Meanwhile, F0 values of unvoiced intervals were set to zero.

Fig. 14 shows the plots of original and reconstructed F0 parameters and the plot of frame-wise F0 error, $e_p(n)$, where $e_p(n) = \hat{p}(n) - p(n)$ with $p(n)$ and $\hat{p}(n)$ are the original and reconstructed F0 parameters, respectively, for the same sentence utterance as in Fig. 10. The RMS F0 error, $\sqrt{E_p}$, where $E_p = \frac{1}{N}\sum_{n=1}^{N} {e_p}^2(n)$, was found to be about 3.6183 Hz.

(Fig. 14 is around here)

4.3.3 Coding gain parameters

The gain contour was re-sampled at 20 ms intervals. Logarithmic quantization was performed using 6 bits for each sampled value. The quantized samples and the spline interpolation were used in the decoder to form the reconstructed gain contour. It should be noted that we did not describe the gain information of speech using the event functions obtained from LEBEL-TD analysis of LSF parameters as in subsections 4.3.1 and 4.3.2. This is due to the fact that the gain parameters are quickly and frequently changed, which results in a low reconstruction accuracy if the same method described in the two previous subsections is applied.

4.4 Bit allocation

Table 3 shows the bit allocation for the proposed speech coding method. An example of bit-rate contour for a male/Japanese utterance is shown in Fig. 15. Note that the average number of events per second, i.e. the event rate, was set as 25 events/sec, resulting in the average algorithmic delay of 55 ms. The larger the event rate, the better the speech quality, however at the cost of increasing the bit rate required for encoding speech. We can control the peak bit rate by, for example, setting the minimum length between the locations of two consecutive events.

(Table 3 is around here)

(Fig. 15 is around here)

4.5 Subjective test

In order to evaluate the performance of the proposed speech coding method, the quality of the reconstructed speech was compared to that of other low bit rate speech coders such as the 4.8 kbps FS-1016 CELP (Campbell et al., 1991) and 2.4 kbps FS-1015 LPC-10E coders (Campbell and Tremain, 1991). By this we show that the proposed speech coding method can achieve good-quality speech with less than 2 kbps.

A subjective test was carried out using the Scheffe's method of paired comparison (Scheffe, 1952). Six graduate students known to have normal hearing ability were recruited for the listening experiment. Each listener was asked to grade from -2 to 2 the degradation perceived in speech quality when comparing the second stimulus to the first, in each pair. The Japanese speech dataset used in Section 3 and an English speech dataset collected from the TIMIT speech corpus (Garofolo et al., 1993), which consists of 192 sentence utterances spoken by 24 speakers (18 male & 6 female), were selected as the training data for the proposed speech coder. They were re-sampled at 8 kHz sampling frequency and STRAIGHT analyzed using the frame shift of 1 ms. LSF transformation was then performed and the resulting 32^{nd} order LSF parameters were TD analyzed by using the LEBEL-TD method. It should be noted that the higher order of LSFs, the better quality of encoded speech. However, it was empirically perceived that a considerable improvement of speech quality is not achieved when the order of LSFs exceeds 32.

Eight phoneme balanced sentences, which are out of the training set, uttered by 4 English (2 male & 2 female) and 4 Japanese (2 male & 2 female) speakers were used as the test data. Namely, the test data comprises 4 male and 4 female utterances. Stimuli were synthesized by using the following coders: 4.8 kbps FS-1016 CELP, 2.4 kbps FS-1015 LPC-10E, and the proposed 1.8 kbps speech coder. Also, 16 other stimuli were STRAIGHT synthesized using the unquantized speech parameters obtained from STRAIGHT analysis & LSF transformation (STRAIGHT-LSF) as well as STRAIGHT analysis, LSF transformation & LEBEL-TD analysis (STRAIGHT-LSF & LEBEL-TD) of the above 8 utterances.

(Fig. 16 is around here)

Results of the listening experiment are shown in Fig. 16. It can be seen from this figure that the quality of the reconstructed speech obtained from the proposed speech coder is comparable to that of the 4.8 kbps FS-1016 CELP coder and is much better than that of the 2.4 kbps FS-1015 LPC-10E coder. This justifies the usefulness of the proposed LEBEL-TD algorithm when being applied to coding speech at low-bit rates.

5 Conclusion

In this paper we have presented a new algorithm for temporal decomposition of speech. The proposed LEBEL-TD method uses the limited error criterion for initially estimating the event locations, and then further refines them using the local optimization strategy. This method achieves results comparable to other TD methods such as S^{2}BEL-TD and MRTD while requiring less algorithmic delay and less computational cost. Moreover, the buffering technique used for continuous speech analysis has been well developed and the stability of the corresponding LPC synthesis filter after spectral transformation performed by LEBEL-TD has been completely ensured. It is shown that the temporal pattern of the speech excitation parameters can also be well described using the LEBEL-TD technique.

We have also described a method for variable-rate speech coding based on STRAIGHT using LEBEL-TD. For encoding spectral information of speech, LEBEL-TD based vector quantization was used. Other speech parameters were quantized by scalar quantization. As a result, a variable-rate speech coder operating at rates around 1.8 kbps was produced. The quality of the reconstructed speech is comparable to that of the 4.8 kbps FS-1016 CELP coder according to the listening experiment. It was shown that the proposed speech coding method can produce good quality speech with less than 2 kbps.

Acknowledgements

This work was partially supported by CREST (Core Research for Evolutional Science and Technology) of JST (Japan Science and Technology Corporation). The authors thank anonymous reviewers for their suggestions and comments that improved the presentation of the paper.

References

Atal, B.S., 1983. Efficient coding of LPC parameters by temporal decomposition. In: Proc. ICASSP, pp. 81-84.

Athaudage, C.N., Brabley, A.B., Lech, M., 1999. Optimization of a temporal decomposition model of speech. In: Proc. International Symposium on Signal Processing and Its Applications, Australia, pp. 471-474.

Campbell, J.P.Jr. and Tremain, T.E., 1986. "Voiced/Unvoiced Classification of Speech with Applications to the U.S. Government LPC-10E Algorithm," In: Proc. ICASSP, pp. 473-476.

Campbell, J.P.Jr., Tremain,T.E., Welch, V.C., 1991. The Federal Standard 1016 4800 bps CELP Voice Coder. Digital Signal Processing, Vol. 1, No. 3, pp. 145-155.

Dix, P.J., Bloothooft, G., 1994. A breakpoint analysis procedure based on temporal decomposition. IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 1, pp. 9-17.

Fallside, F., Woods, W.A. (Eds.), 1985. Computer Speech Processing. Prentice-Hall, New York.

Ghaemmaghami, S., Deriche, M., Boashash, B., 1997. Comparative study of different parameters for temporal decomposition based speech coding. In: Proc. ICASSP, pp. 1703-1706.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., 1993. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM, NIST.

Itakura, F., 1975. Line spectrum representation of linear predictive coefficients of speech signals. Journal of the Acoustical Society of America, Vol. 57, p. S35.

Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication, Vol. 27, No. 3-4, pp. 187-207.

Kim, S.J., Oh, Y.H., 1999. Efficient quantization method for LSF parameters based on restricted temporal decomposition. Electronics Letters, Vol. 35, No. 12, pp. 962-964.

Linde,Y., Buzo, A., Gray, R.M., 1980. An algorithm for vector quantiser design. IEEE Trans. on Communication, Vol. 28, pp. 84-95.

Nandasena, A.C.R., Nguyen, P.C., Akagi, M., 2001. Spectral stability based event localizing temporal decomposition. Computer Speech and Language, Vol. 15, No. 4, pp. 381-401.

Nguyen, P.C., Akagi, M., 2002. Improvement of the restricted temporal decomposition method for line spectral frequency parameters. In: Proc. ICASSP, pp. 265-268.

Nguyen, P.C., Akagi, M., 2002. Limited error based event localizing temporal decomposition. In: Proc. EUSIPCO, pp. 239-242.

Nguyen, P.C., Ochi, T., Akagi, M., 2002. Coding speech at very low rates using STRAIGHT and temporal decomposition. In: Proc. ICSLP, pp. 1849-1852.

Nguyen, P.C., Akagi, M., 2002. Variable rate speech coding using STRAIGHT and temporal decomposition. In: Proc. IEEE Speech Coding Workshop, pp. 26-28.

Niranjan, M., Fallside, F., 1989. Temporal decomposition: a framework for enhanced speech recognition. In: Proc. ICASSP, pp. 655-658.

Paliwal, K.K., 1995. Interpolation properties of linear prediction parametric representations. In: Proc. Eurospeech, pp. 1029-1032.

Paliwal, K.K., Atal, B.S., 1993. Efficient vector quantization of LPC parameters at 24 bits/frame. IEEE Transactions on Speech and Audio Processing, Vol. 1, No. 1, pp. 3-14.

Scheffe, H., 1952. An analysis of variance for paired comparisons. Journal of the American Statistical Association, Vol. 47, pp. 381-400.

Shiraki, Y., Honda, M., 1991. Extraction of temporal pattern of spectral sequence based on minimum distortion criterion. In: Proc. Autumn Meeting of the Acoustical Society of Japan, pp. 233-234 (in Japanese).

Van Dijk-Kappers, A.M.L., Marcus, S.M., 1989. Temporal decomposition of speech. Speech Communication, Vol. 8, pp. 125-135.

Table 1
Event rate, average LSD, and percentage number of outlier frames obtained from the LEBEL-TD, S^{2}BEL-TD, RTD and MRTD methods. The spectral parameter is LSF. Speech dataset consists of 250 sentence utterances spoken by 10 speakers (5 male & 5 female) of the ATR Japanese speech database.

Method	Event rate	Avg. LSD	2-4 dB	> 4 dB
LEBEL-TD	19.996	1.5125 dB	32.52%	0.07%
S^{2}BEL-TD	19.455	1.4643 dB	18.48%	0.94%
RTD	20.163	1.5629 dB	22.97%	0.96%
MRTD	20.163	1.5681 dB	23.15%	0.98%

Table 2
Event rate, average LSD, and percentage number of outlier frames obtained from the LEBEL-TD method for some ε. The spectral parameter is LSF. Speech dataset consists of 250 sentence utterances spoken by 10 speakers (5 male & 5 female) of the ATR Japanese speech database.

ε	Event rate	Avg. LSD	2-4 dB	$>$ 4 dB
0.072	15.059	1.9220 dB	48.52%	1.60%
0.065	16.051	1.8255 dB	45.54%	0.89%
0.058	17.156	1.7270 dB	41.91%	0.46%
0.053	18.107	1.6491 dB	38.69%	0.23%
0.049	18.999	1.5802 dB	35.64%	0.13%
0.045	19.996	1.5125 dB	32.52%	0.07%
0.041	21.124	1.4400 dB	29.02%	0.04%
0.038	22.117	1.3833 dB	26.21%	0.02%
0.0355	23.050	1.3331 dB	23.75%	0.014%
0.033	24.106	1.2795 dB	20.96%	0.01%
0.031	25.028	1.2336 dB	18.69%	0.00%

Table 3
Bit allocation for the proposed speech coder.

Parameter	Proposed Speech Coder
Event vector	32 bits (8+8+8+8)
Event function	7 bits
Event location	8 bits
F0 target	6 bits
Noise ratio target	6 bits
Subtotal A (sum × event rate)	1475 bps
Gain	300 bps
Lengths of voiced and unvoiced segments	36 bps
Maximum amplitude of input speech	5 bps
Subtotal B	341 bps
Total (A+B)	1816 bps

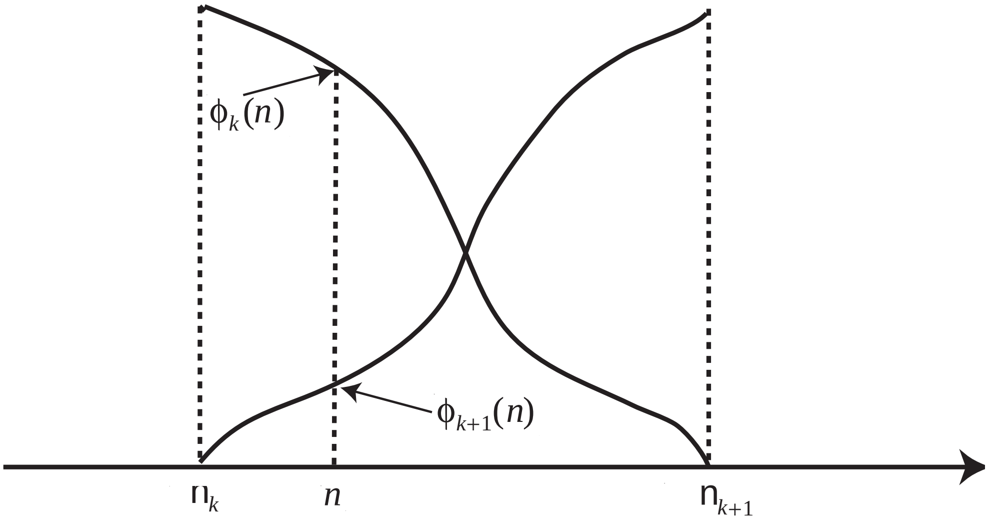


Fig. 1. Example of two adjacent event functions in the second order TD model.

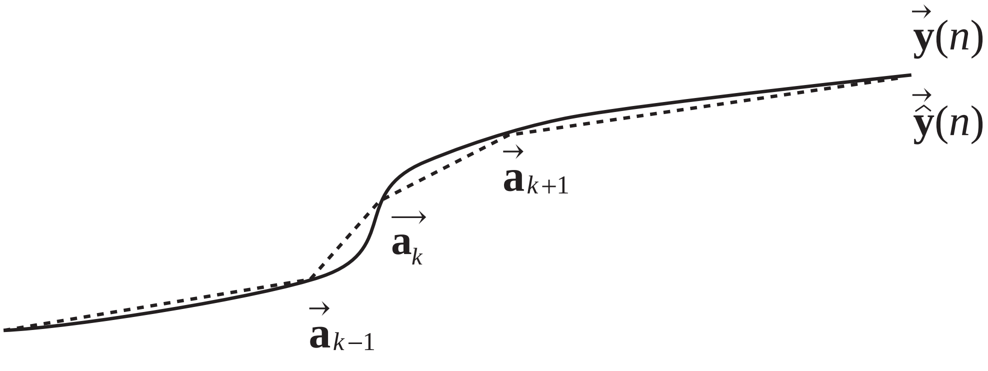
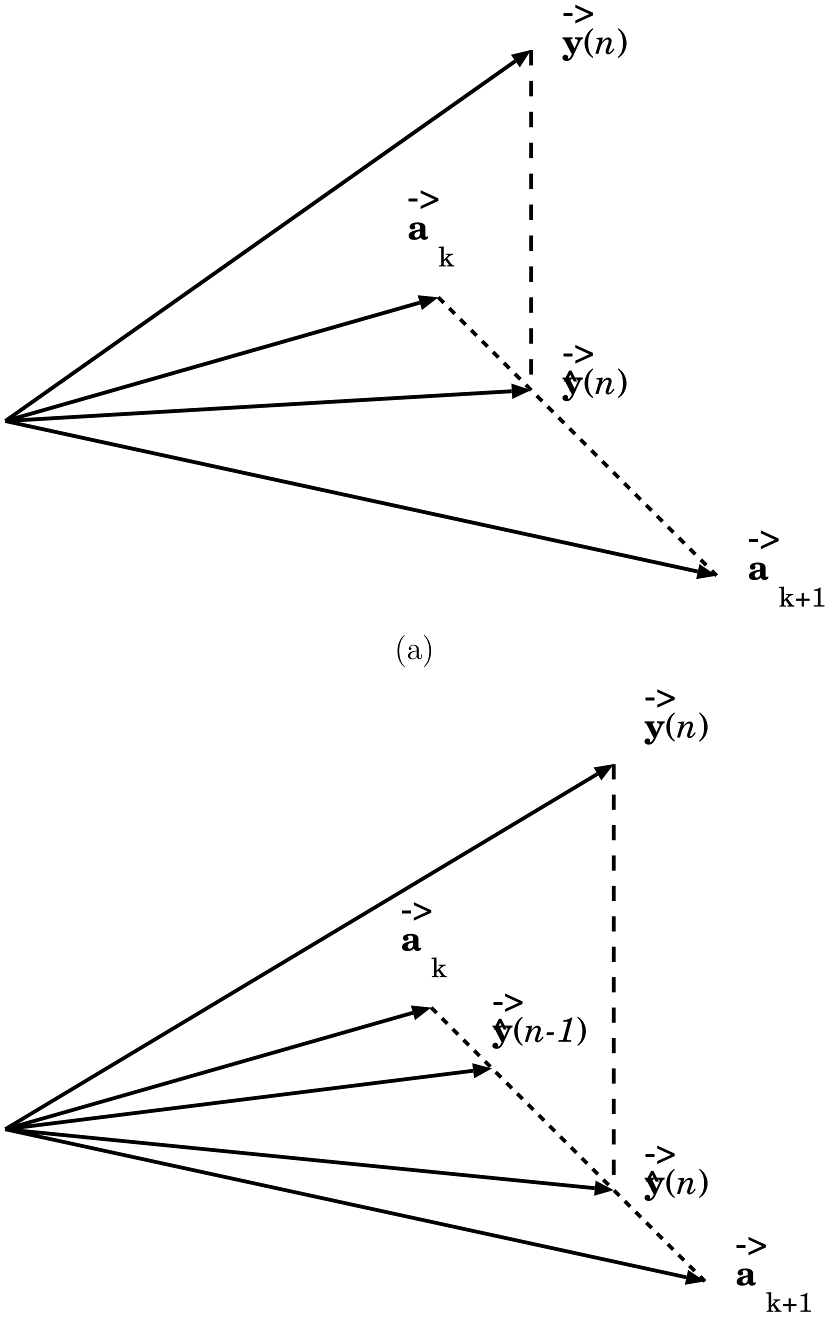


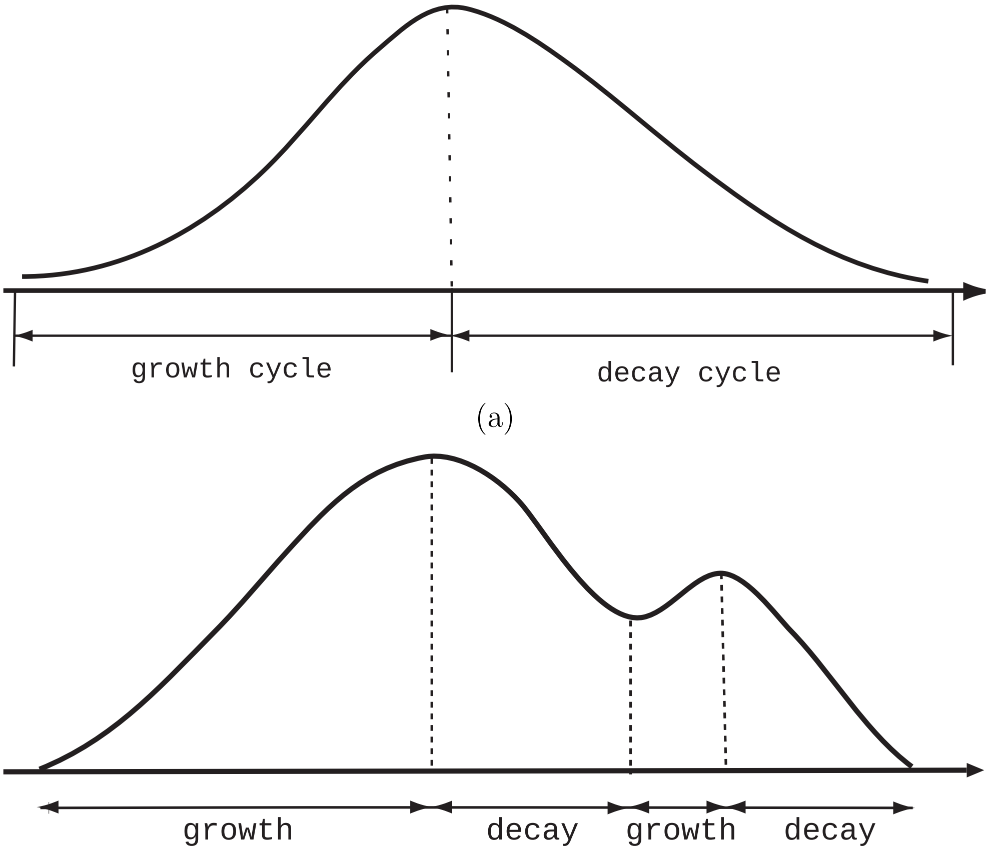
Fig. 2. The path in parameter space described by the sequence of spectral parameters $\mathbf{y}(n)$ is approximated by means of straight line segments between breakpoints.



(a)

(b)

Fig. 3. Determination of the event functions in the transition interval $[n_k, n_{k+1}]$. The point of the line segment between $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ (a), between $\hat{\mathbf{y}}(n-1)$ and $\mathbf{a}_{k+1}$ (b) with minimum distance from $\mathbf{y}(n)$ is taken as the best approximation.



(a)

(b)

Fig. 4. Examples of a well-shaped event function (a) and an ill-shaped event function (b).

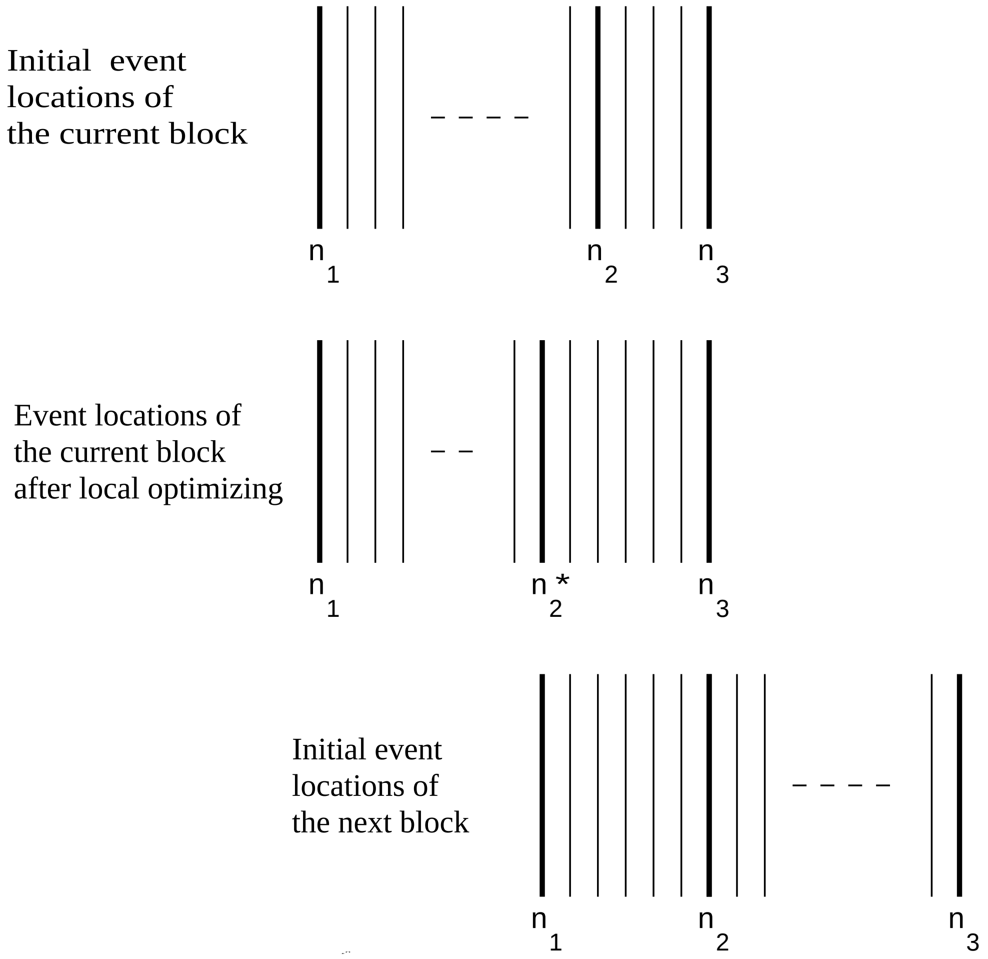

Initial event locations of the current block
n
1
n
2
n
3
Event locations of the current block after local optimizing
n
1
n *
2
n
3
Initial event locations of the next block
n
1
n
2
n
3

Fig. 5. Buffering technique for LEBEL-TD

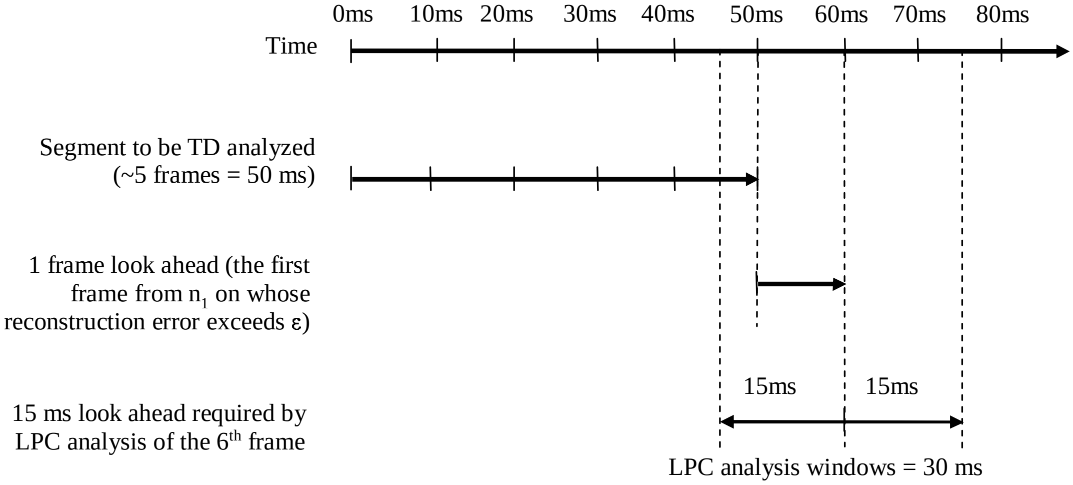

0ms
10ms
20ms
30ms
40ms
50ms
60ms
70ms
80ms
Time
Segment to be TD analyzed (~5 frames = 50 ms)
1 frame look ahead (the first frame from n_1 on whose reconstruction error exceeds ε)
15ms
15ms
15 ms look ahead required by LPC analysis of the 6th frame
LPC analysis windows = 30 ms

Fig. 6. Algorithmic delay for LEBEL-TD

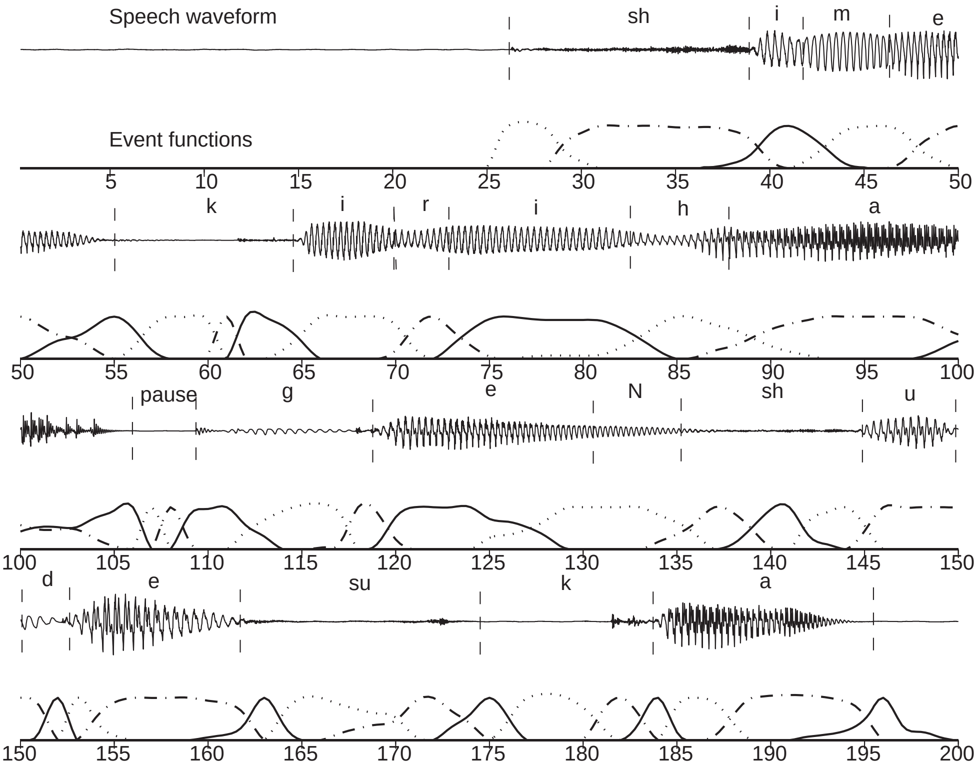


Fig. 7. Plot of the event functions obtained from the LEBEL-TD method for the female/Japanese sentence utterance 'shimekiri ha geNshu desu ka.' The speech waveform is also shown together with the phonetic transcription for reference. The numerals indicate the frame numbers.

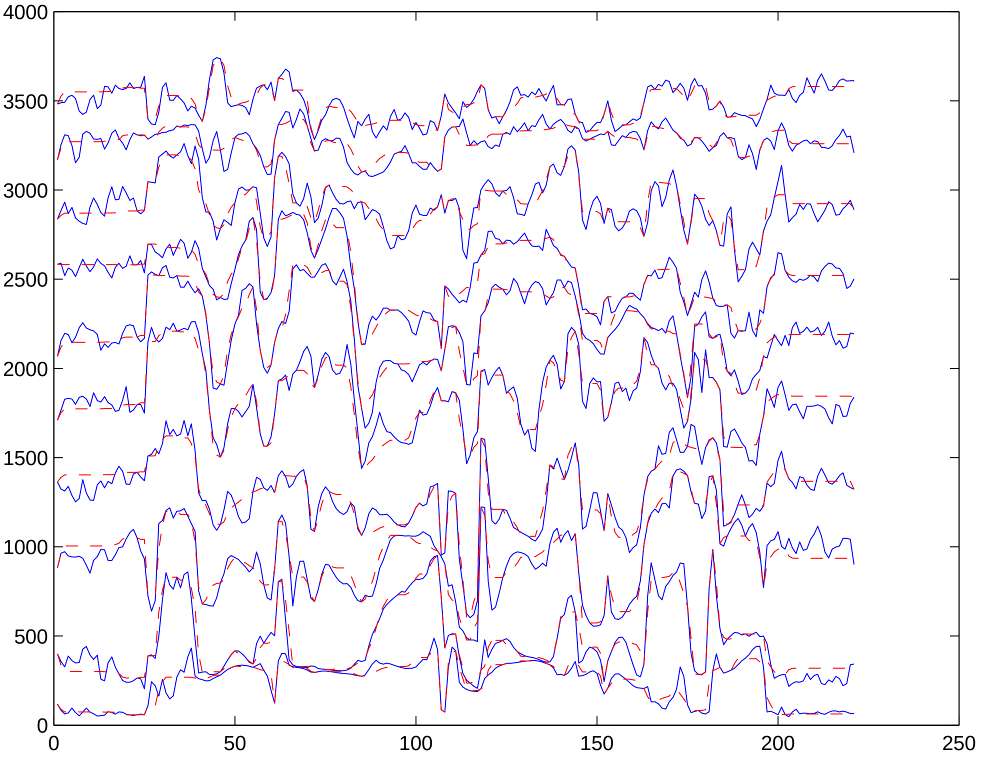


Fig. 8. Plots of the original and reconstructed LSF parameters obtained from the LEBEL-TD method for the female/Japanese speech utterance “*shimekiri ha geNshu desu ka*.” The solid line indicates the original LSF parameter vector trajectory and the dashed line indicates the reconstructed LSF parameter vector trajectory. The average log spectral distortion was found to be 1.6276 dB.

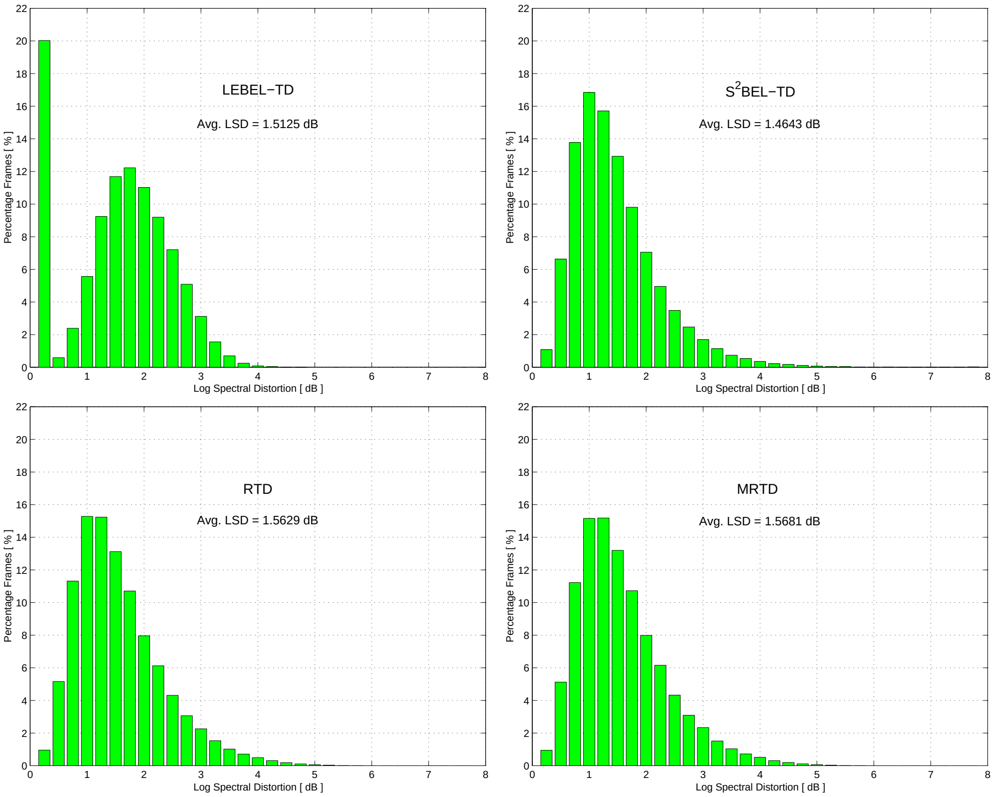


Fig. 9. Distribution of the Log Spectral Distortion (LSD) between the original and reconstructed LSF parameters in the form of histograms. Top left: LSD histogram for LEBEL-TD. Top right: LSD histogram for S^{2}BEL-TD. Bottom left: LSD histogram for RTD. Bottom right: LSD histogram for MRTD. Speech dataset consists of 250 sentence utterances spoken by 10 speakers (5 male & 5 female) of the ATR Japanese speech database.

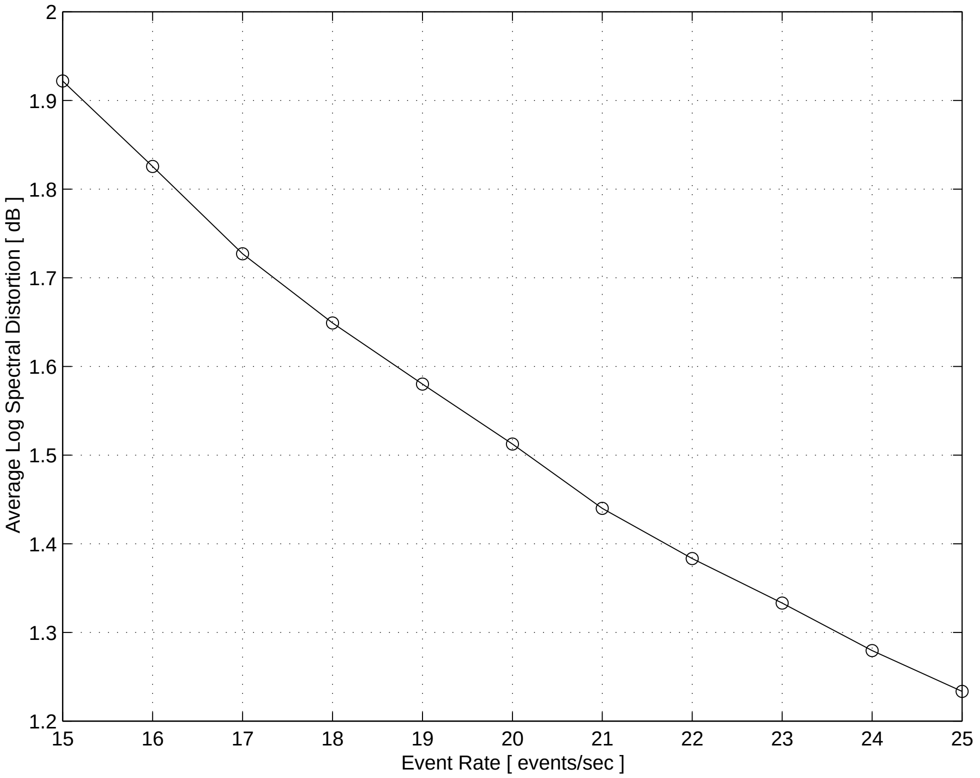


Fig. 10. Average log spectral distortion (dB) versus the event rate (events/sec).

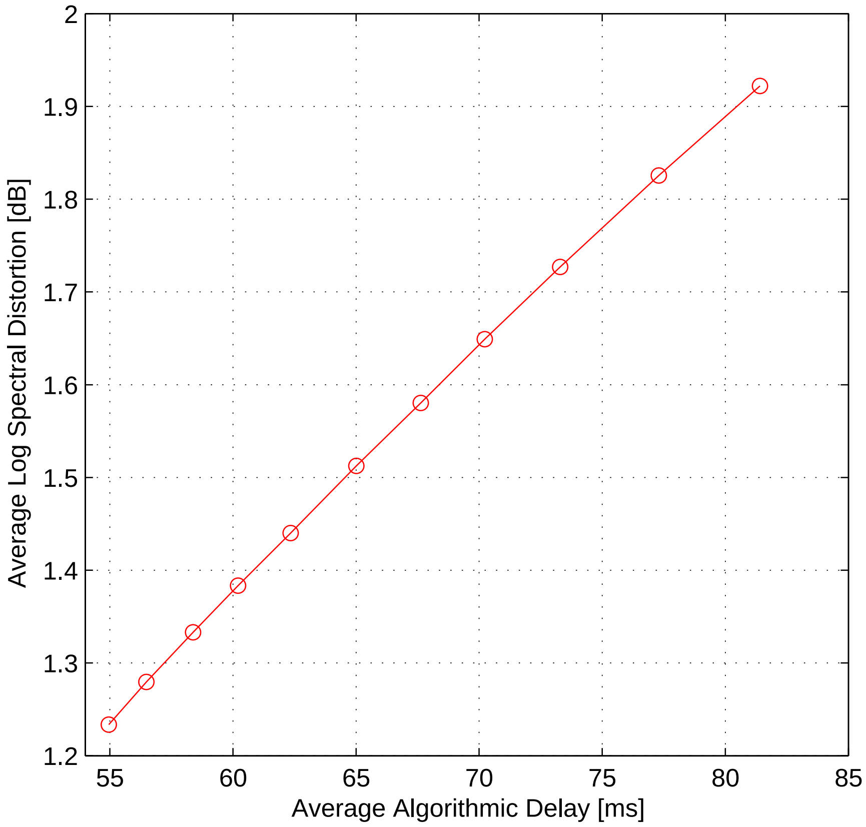


Fig. 11. Average log spectral distortion (dB) versus the average algorithmic delay (ms).

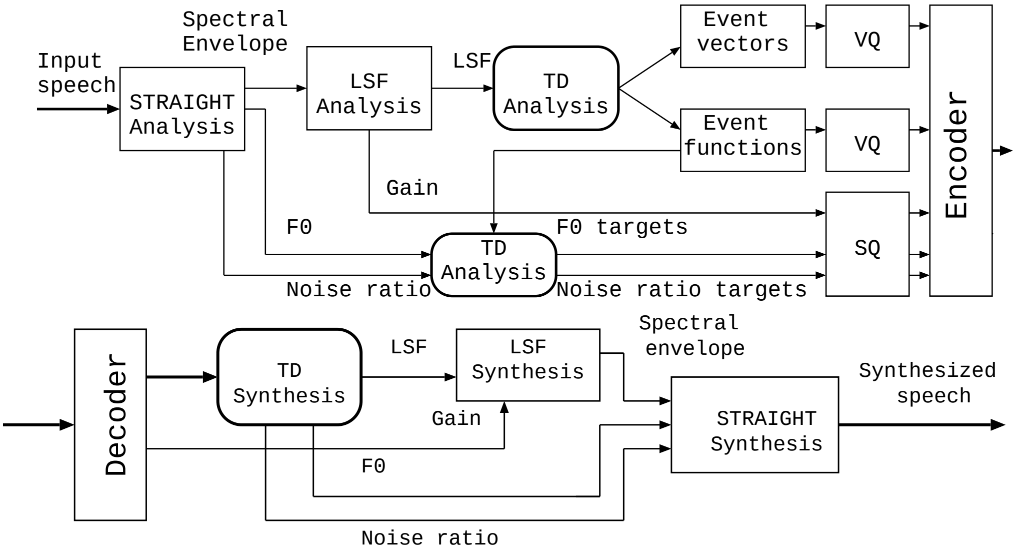


Fig. 12. Proposed speech encoder and decoder block diagrams (top: encoder, bottom: decoder).

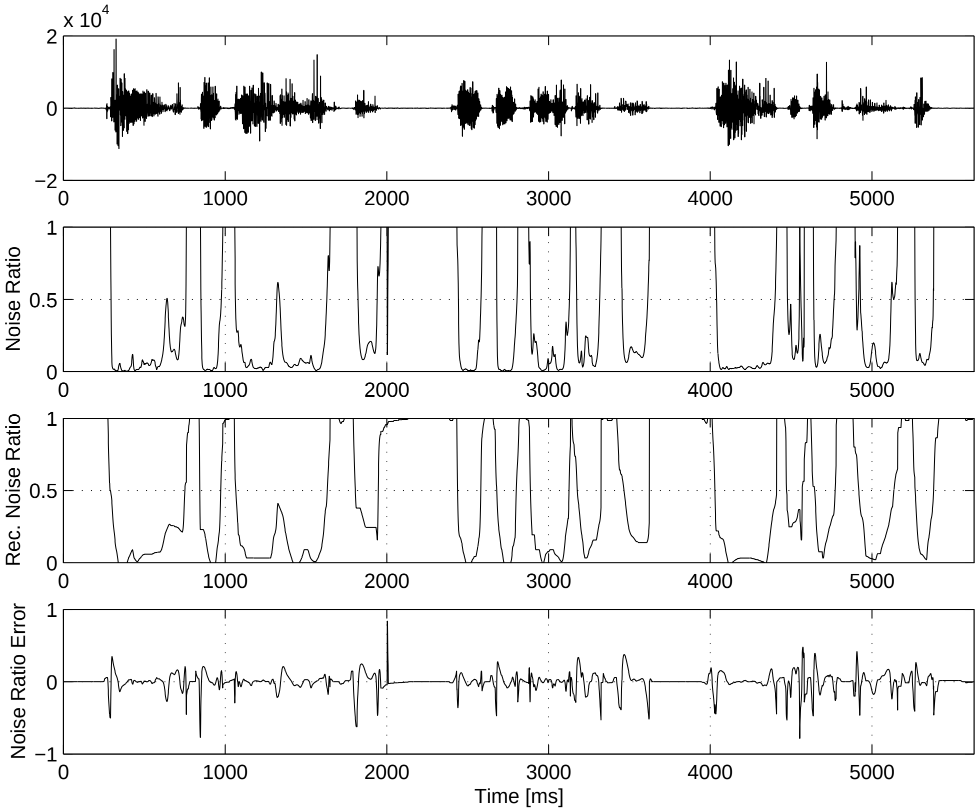


Fig. 13. Original noise ratio parameters, $i(n)$, reconstructed noise ratio parameters, $\hat{i}(n)$, and frame-wise noise ratio error, $e_i(n) = \hat{i}(n) - i(n)$, for the sentence utterance 'kaigi ni happyou surunodeha nakute choukou surudake dato, hiyou ha ikura kakari masu ka,' of the ATR Japanese speech database. The RMS noise ratio error is 0.1166. The speech waveform is also shown together for reference.

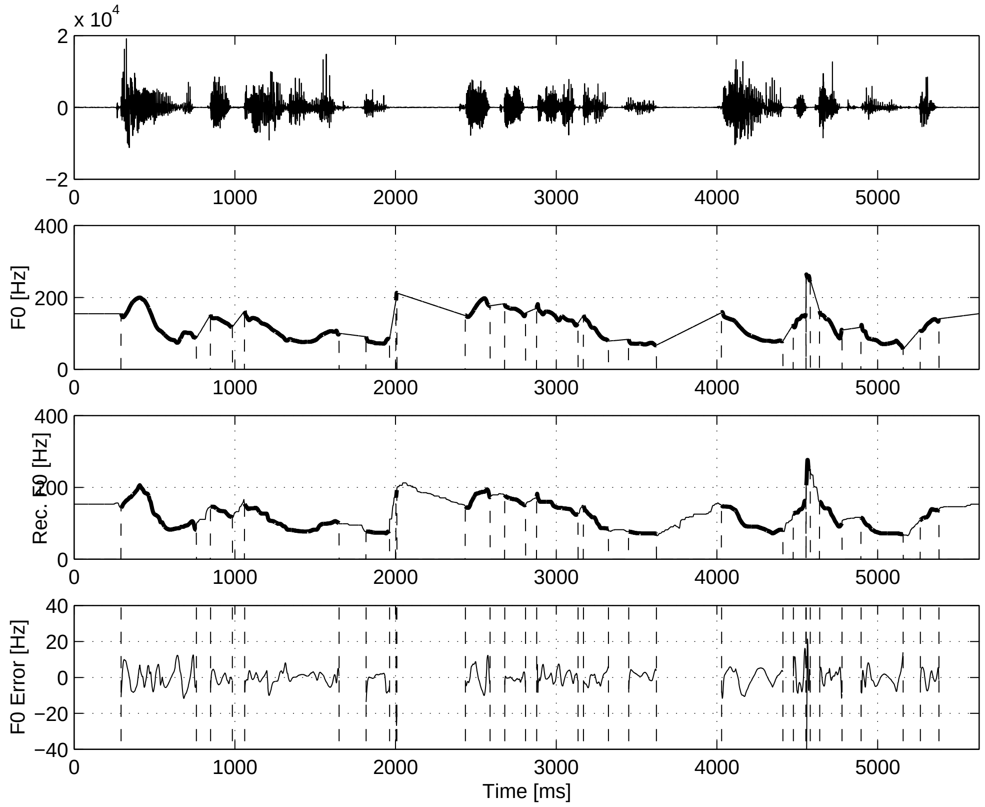


Fig. 14. Original F0 parameters, $p(n)$, reconstructed F0 parameters, $\hat{p}(n)$, and frame-wise F0 error, $e_p(n) = \hat{p}(n) - p(n)$, for the sentence utterance 'kaigi ni happyou surunodeha nakute choukou surudake dato, hiyou ha ikura kakari masu ka,' of the ATR Japanese speech database. F0 error is shown only for the voiced segments of the utterance. The RMS F0 error is 3.6183 Hz. The speech waveform is also shown together for reference.

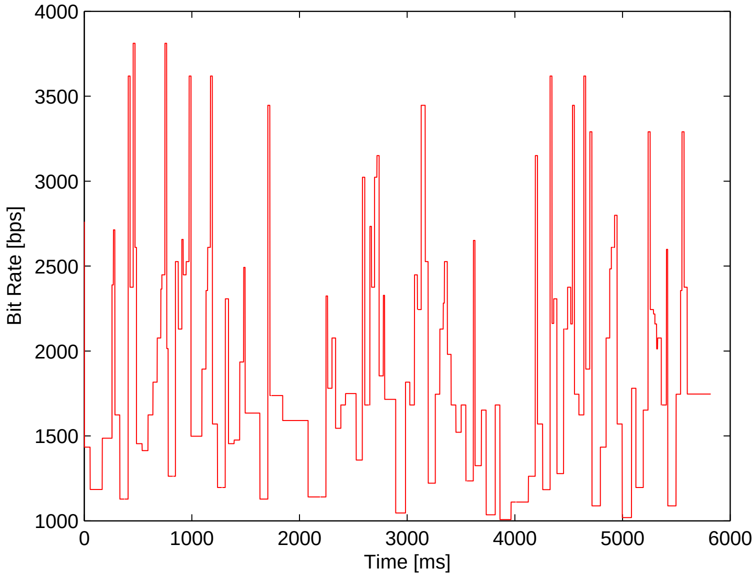


Fig. 15. Bit-rate contour for a male/Japanese sentence utterance 'konkai no kokusaikaigi ha tuuyaku denwa ni kansuru naiyou wo subete fukunde imasu.'

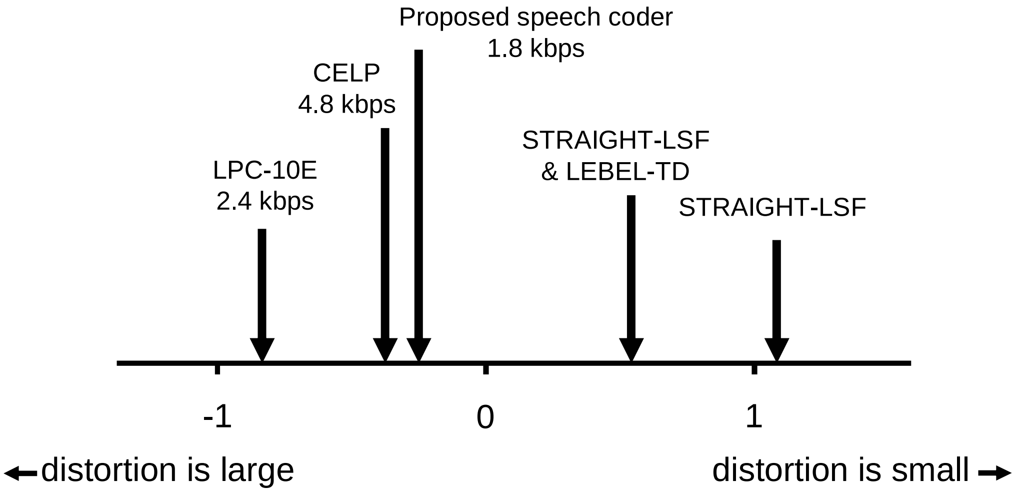


Fig. 16. Results of the listening experiment.