REPRESENTATION OF SPECTRAL ENVELOPE WITH WARPED FREQUENCY RESOLUTION FOR AUDIO CODER

R. Sugiura[1], *Y. Kamamoto*[2], *N. Harada*[2], *H. Kameoka*[2], *T. Moriya*[2]

[1]Graduate School of Information Science and Technology, The University of Tokyo
[2]NTT Communication Science Labs., Nippon Telegraph and Telephone Corp.

ABSTRACT

We have devised a method for representing frequency spectral envelopes with warped frequency resolution based on sparse non-negative matrices aiming at its use for frequency domain audio coding. With optimally prepared matrices, we can selectively control the resolution of spectral envelopes and enhance the coding efficiency. We show that the devised method can enhance the subjective quality of the state-of-the-art wide-band coder at 16 kbit/s at a cost of minor additional complexity. The method is therefore, expected to be useful for low-bit-rate and low-delay audio coder for mobile communications.

Index Terms— audio coding, signal processing, frequency warping, non-negative matrix, TCX

1. INTRODUCTION

For years, speech coders have been developed for use in voice communication tools such as mobile phones. However, there is a demand for higher quality not only in speech, but also in the other audio signals such as music.

3GPP Extended Adaptive Multi-Rate WideBand (AMR-WB+) and MPEG-D Unified Speech and Audio Coding (USAC) [1,2] are the state-of-the-art speech and audio coders. Both have at least two different modes and switch from one to the other depending on the input signals. Voice signals are coded in the time domain, and the other audio signals are coded in the frequency domain by Transform Coded eXitation (TCX). The goal of the present work is to design a low-bit-rate audio coder with higher quality, with lower algorithmic delay than AMR-WB+ or USAC so that the coder can be used in mobile communications. Here, to achieve higher quality in the audio coder, we discuss modifying the TCX coder.

The efficiency of TCX is highly dependent on the parameters representing frequency spectral envelopes of the inputs. In this paper, we introduce a model using frequency warping for a more efficient representation of spectral envelopes with a similar motivation such as in [3–6]. To achieve low computational complexity for the warping, we construct sparse matrices with non-negative elements that approximate warping and inverse warping. In addition, perceptual weighting for the envelopes of this model is also considered.

In section 2, we outline the TCX coder. Then, in section 3, the model of spectral envelopes is introduced. Finally, in section 4, the model is evaluated by quantitative and subjective assessments.

2. FREQUENCY DOMAIN AUDIO CODING

2.1. TCX coder

The state-of-the-art TCX coder uses Modified Discrete Cosine Transform (MDCT) [7] and is used in USAC. Fig. 1 describes the TCX process. The coder quantizes and codes two kinds of information: linear prediction (LP) coefficients, which represent the spectral envelopes, and residual spectra, the frequency spectra divided by their envelopes. LP coefficients are first transformed into line spectrum pairs (LSP), which are robust for interpolation and quantization, and then vector quantized. On the other hand, residual spectra are compressed by entropy coding after they are scalar quantized for each frequency bin.

Envelopes are perceptually weighted when divided into the spectra. Normally, envelopes are calculated from LP coefficients by using an all-pole filter:

$$H_k = 1/|1 + \sum_n a_n e^{-j\frac{2\pi k}{N}n}|,\ (k = 0, \cdots, N-1) \quad (1)$$

where N and a_n are the length of the envelopes and the coefficient of the nth order, respectively. To make the quantization noise in the spectra less annoying, the envelopes are smoothed by the weighting defined as

$$\tilde{H}_k = 1/|1 + \sum_n a_n \gamma^n e^{-j\frac{2\pi k}{N}n}|,\ (0 < \gamma < 1). \quad (2)$$

This weighting shapes the quantization noise into approximately $\tilde{H}_k/H_k$, eventually resulting in smaller distortion in the spectral peaks, which is more important than the spectral valleys for human perception. It is experimentally known that the perceptual weighting works well when 0.92 is chosen for the value of the parameter γ.

Moreover, the envelopes can improve the performance of the entropy coding. The values of the residual spectra can be roughly estimated from the values of their envelopes: the higher the values of the envelopes are, the more likely the values of the residual spectra are to be high.

2.2. Limitation of linear prediction

As stated above, envelopes can shape the quantization noise in the spectra to realize a perceptually efficient quantization and can also estimate the values of the residual spectra. However, this noise shaping approximation and value estimation stand

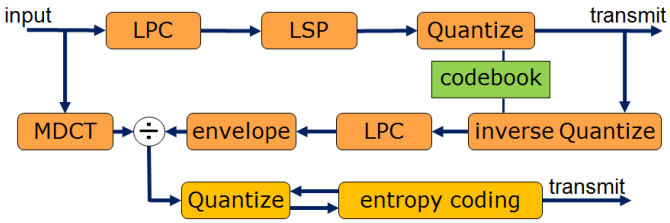


Fig. 1. Outline of TCX encoder.

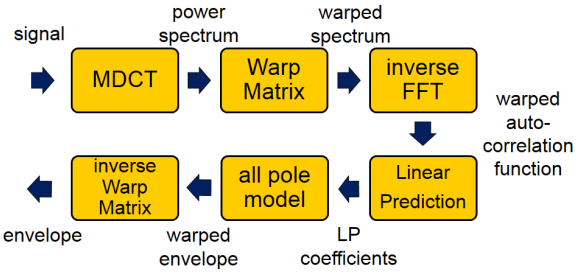


Fig. 2. Flowchart for extracting resolution warped envelopes

only if the spectral envelopes properly represent the shapes of their spectra. Generally, the envelopes extracted by linear prediction have about (length of the signal) /(order of LP) frequency resolution and this resolution is uniform over the frequency axis. This limitation on the resolution makes, in some cases, the envelopes fail to represent their spectra, which leads to an unexpected distribution of the quantization noise and estimation in the entropy coding. Indeed, this failure can be avoided by increasing the order of the prediction, but this also increases the parameters that must to be transmitted. In fact, most natural sounds have a lot of power in the lower frequencies so that the information content of quantized signals tends to be biased to the lower band. Therefore, instead of transmitting more parameters, we modify the model of envelopes to warp the frequency resolution into, for example, a Mel-frequency scale.

3. DESIGN OF THE ENVELOPE MODEL

3.1. Extracting envelopes using frequency warping

A previous work on modeling spectral envelopes with warped resolution used an all-path filter to modify the all-pole model eq. (1) [4]. This modified method is called Mel Linear Predictive Coding (Mel-LPC) and is a special case of Mel generalized cepstral analysis [5], of which use was considered in a time-domain speech coder [6]. However, estimating the parameters of the model from inputs and calculating envelopes both requires much more computational complexity than linear prediction. In addition, this filter has only one parameter for tuning the warping and is lacking in flexibility.

To reduce computational costs, the model introduced here uses sparse non-negative matrices for approximating frequency warping. Fig. 2 shows the processes for extracting resolution warped envelopes. First, power spectra are frequency warped by a warping matrix and Fourier transformed. Then, the transformed spectra are used as a pseudo-autocorrelation function to estimate parameters by linear predictive analysis. Finally, envelopes are calculated by substituting the parameters into the all-pole model eq. (1), followed by inverse warping with another matrix operation. Since we use the linear predictive analysis, the parameters are explicitly found and the stability can be easily guaranteed.

3.2. Approximation of warping by a matrix operation

With discrete signals, frequency warping is an irreversible operation, which involves a risk of unexpected transformations when the signals are warped and inversely warped by simple sinc interpolation. To reduce the risk, it is reasonable to find the warping and inverse warping matrices in advance by optimizing for a training data set. For the measurement of this optimization, Itakura-Saito distance is selected, as in linear prediction [8]. The objective function to minimize for optimizing the warping matrix W is defined as

$$\sum_{i,j}\left(\frac{Y_{ij}}{\sum_k W_{ik}X_{kj}} - \log\frac{Y_{ij}}{\sum_k W_{ik}X_{kj}} - 1\right) \tag{3}$$

where X and Y respectively stands for the power spectra of the training data set and the spectra warped by sinc interpolation. Subscript i and k indicate the number of frequency bin, and j indicates the number of frame. The function above means an Itakura-Saito distance of WX from Y. Since the spectra warped by W must still be power spectra, the elements of W have to be found in a non-negative condition. This minimization problem cannot be solved explicitly, which leads to a need for an implicit approach such as the auxiliary function method [9] as follows.

First of all, we consider only the terms in the objective function eq. (3) related to W and minimizing it:

$$\sum_{i,j}\frac{Y_{ij}}{\sum_k W_{ik}X_{kj}} + \sum_{i,j}\log\left(\sum_k W_{ik}X_{kj}\right) \equiv L(W). \tag{4}$$

Since the function $f(x) = \frac{1}{x}$ is convex in $x > 0$, Jensen's inequality holds as

$$\begin{aligned}\frac{1}{\sum_k W_{ik}X_{kj}} &= \frac{1}{\sum_k \lambda_{ijk}(W_{ik}X_{kj}/\lambda_{ijk})} \\ &\le \sum_k \frac{\lambda_{ijk}}{W_{ik}X_{kj}/\lambda_{ijk}} \quad \left(\lambda_{ijk} \ge 0,\ \sum_k \lambda_{ijk} = 1\right).\end{aligned} \tag{5}$$

In addition, the concavity of the logarithmic function leads to

$$\log\left(\sum_k W_{ik}X_{kj}\right) \le \log\phi_{ij} + \frac{\sum_k W_{ik}X_{kj}}{\phi_{ij}},\ (\phi_{ij} > 0). \tag{6}$$

Using both inequalities (5) and (6), the upper bound can be set to the function $L(W)$ by an auxiliary function as

$$\begin{aligned}L(W) \le &\sum_{i,j} Y_{ij}\sum_k \frac{\lambda_{ijk}^2}{W_{ik}X_{kj}} \\ &+ \sum_{i,j}\left(\log\phi_{ij} + \frac{\sum_k W_{ik}X_{kj}}{\phi_{ij}}\right) \equiv G(W)\end{aligned} \tag{7}$$

provided that the equality is attained if and only if the following equation holds

$$\lambda_{ijk} = \frac{W_{ik}X_{kj}}{\sum_k W_{ik}X_{kj}},\ \ \phi_{ij} = \sum_k W_{ik}X_{kj}. \tag{8}$$

Because of the convexity, the auxiliary function $G(W)$ can be minimized, with λ and ϕ both fixed, at the stationary point of W which is found as

$$\begin{aligned}&\frac{\partial}{\partial W_{mn}} G(W)|_{W=\tilde{W}} \\ &= \sum_j Y_{mj}\lambda_{mjn}^2/X_{nj}\left(-\frac{1}{\tilde{W}_{mn}^2}\right) + \sum_j \frac{X_{nj}}{\phi_{mj}} = 0 \\ &\iff \tilde{W}_{mn} = \sqrt{\frac{\sum_j Y_{mj}\lambda_{mjn}^2/X_{nj}}{\sum_j X_{nj}/\phi_{mj}}}.\end{aligned} \tag{9}$$

At last, by eqs. (8) and (9), λ, ϕ and $\tilde{W}$ are iteratively updated, which makes the objective function decrease monotonically until it ends up in a local optimum. This iteration can be summarized as follows.

Respectively λ, ϕ are updated by $W^{(l)}$, which indicates $\tilde{W}$ in the lth iteration, as

$$\lambda_{ijk} = \frac{W_{ik}^{(l)}X_{kj}}{\sum_k W_{ik}^{(l)}X_{kj}},\ \ \phi_{ij} = \sum_k W_{ik}^{(l)}X_{kj}. \tag{10}$$

Substituting them into eq. (9) leads to $\tilde{W}$ in the $(l+1)$th iteration:

$$W_{mn}^{(l+1)} = \sqrt{\frac{\sum_j Y_{mj}W_{mn}^{(l)2}X_{nj}/\left(\sum_k W_{mk}^{(l)}X_{kj}\right)^2}{\sum_j X_{nj}/\sum_k W_{mk}^{(l)}X_{kj}}}. \tag{11}$$

This results in the following updating rule:

$$W_{mn}^{(l+1)} = W_{mn}^{(l)}\sqrt{\frac{\sum_j Y_{mk}X_{nj}/\hat{Y}_{mj}^2}{\sum_j X_{nj}/\hat{Y}_{mj}}},\ \ \hat{Y} = W^{(l)}X. \tag{12}$$

The updating rule to optimize the inverse warping matrix U can also be derived by using a corresponding method to minimize Itakura-Saito distance of UWX from X.

The updating rules stated above are products of the respective elements and non-negative numbers. Accordingly, the elements of W and U stay positive when the initial values are positive. Moreover, the elements stay zero when the initial values are zero, which enables us to optimize W and U while restricting the number of non-zero elements in the each row. This means the computational complexity of the warping and the inverse warping operations are controllable. In terms of flexibility, this method is capable of designing warping at will by changing the warped spectra Y of the training data.

In the following discussions, we use W and U as respectively the Mel-frequency scale warping matrix and the inverse one, both optimized under the condition of having at most seven elements in each row. Thus, the warping and the inverse warping needs only $7\times$ (frame length) operations of multiplication each, meaning these two operations result in little additional complexity.

3.3. Perceptual weighting for warped envelopes

As mentioned in section 2.1, envelopes extracted from the inputs must be smoothed by the perceptual weighting according to equation (2). However, the weighting needs a modification to apply it to the resolution-warped envelopes stated above. Applying the weighting eq. (2) means smoothing the envelopes with the warped-frequency domain regarded as a linear frequency domain. This mismatch of the domains causes a difference from the expectation in the smoothing result. To distribute the quantization noise according to the shape of $\tilde{H}_k/H_k$, which is known to be less annoying, γ has to be modified for each frequency bin.

Here, we approximate the perceptual weighting as follows. Suppose there are N points in both the linear and warped-frequency domains at equal intervals up to the Nyquist frequency. Then, define k in eq. (2) as an index of frequency in the warped domain and $f(k)$ as an index corresponding to k in the linear domain, provided $f(0) = 0$. Approximation of the weighting can be performed as

$$\tilde{H}_0 = 1/|1 + \sum_n a_n\gamma^n| \tag{13}$$

$$\tilde{H}_k = 1/|1 + \sum_n a_n\gamma^{\frac{f(k)}{k}n}e^{-j\frac{2\pi k}{N}n}|,\ \ (k = 1, \cdots, N-1).$$

Additionally, $f(k)$ stated above can be approximated by

$$(f(0), \ldots, f(N-1))^T \simeq U(0, \ldots, N-1)^T \tag{14}$$

and the smoothing also works in this case.

4. EVALUATIONS

4.1. Fidelity of the envelopes

To evaluate the method described above, we compared envelopes extracted in different ways. For each frame in the input signals, envelopes with resolution warped into the Mel-frequency scale were extracted respectively by Mel-LPC method and by the method proposed in an earlier section, using either the optimized matrix or sinc interpolation for frequency warping. The warp matrix was optimized using training data set containing six minutes of audio and speech signals other than the following test items. Frequency warping by sinc interpolation simply warped the frequency according to the Mel-frequency function and inverse warped according to the inverse function.

Fig. 3 is the clear example for the comparison of the envelopes extracted from one frame by linear prediction and the method using the optimized matrix. The spectrum is presented in linear-amplitude domain since it is quantized in this domain when coded. It is obvious that the resolution of the envelope was warped: the fidelity of the envelope improved in the lower band, which is more important for the coding, at the cost of the higher band.

We also performed a quantitative evaluation. In each frame of the test data, the improvements from linear prediction in Itakura-Saito distance were compared (Fig. 4). The three methods had similar influence on the fidelity of the envelopes. The comparison between the two warping methods reveals the contribution of the optimization to preventing unexpected transformations in the higher band. Therefore it

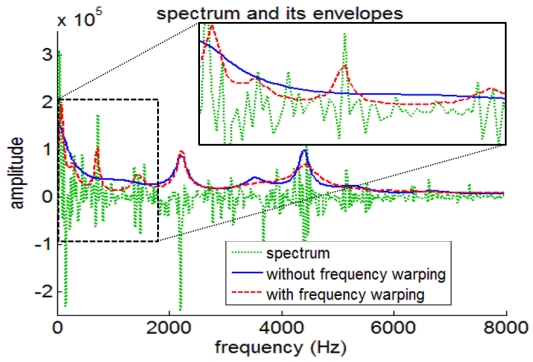


Fig. 3. Comparison of the envelopes. The green dotted line shows the spectrum of the input. The blue solid line and the red dashed line respectively indicate its envelopes extracted using linear prediction and the proposed method. The LP order of the envelopes was 16.

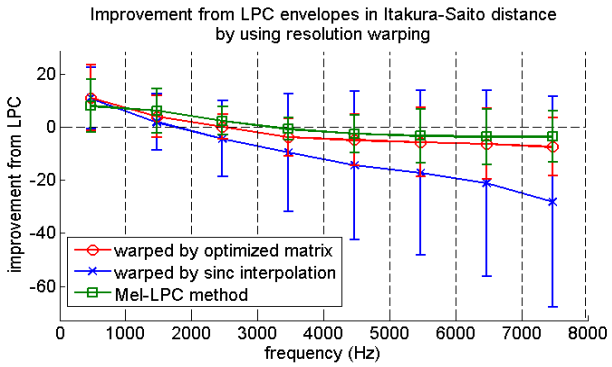


Fig. 4. The average and standard deviation of the improvements from linear prediction in each band. In the band where the improvements are positive, the envelopes were closer to their spectra in Itakura-Saito distance compared with the envelopes extracted by linear prediction. For the test data, 30 second segments from 15 music files in the RWC Music Database were randomly selected. 16-kHz sampling rate, using 256 samples for each frame, and the LP order of the envelopes was 16. We used the Mel-LPC method in Speech Signal Processing Toolkit [10].

can be said that the optimization succeeded in making an appropriate interpolation for the inverse warping. On the other hand, compared with Mel-LPC method, the proposed method showed almost the same performance.

4.2. Perceptual weighting

Secondly, we compared the perceptual weighting for envelopes with warped resolution. Generally, smoothing by the weighting eq. (2) has more influence on the steep peaks than on the gentle slopes of the envelopes. However, as Fig. 5(a), for example, shows, the weighting failed to smooth the resolution-warped envelopes, with peaks remaining in the lower band. This problem arose because the weighting eq. (2) smoothed the envelope in the warped frequency domain instead of the linear domain. By modifying the weighting according to eq. (13), as shown in Fig. 5(b), we were able to smooth the envelope appropriately in accordance with the warping. The local roughness in the smoothed envelopes were caused by the approximation errors in the warping and inverse warping.

4.3. Subjective evaluation of sound quality

Finally, for subjective evaluation, we made a coder based on TCX using the warped envelopes. Fig. 6 shows the outline of the process. The residual spectra of inputs were Golomb-Rice coded [11] after scalar quantization, with the optimal Rice parameter estimated from their envelopes. Thus, the fidelity of the envelopes increases the efficiency of the entropy coding as well as the efficiency of the distribution of the quantized noise. Additionally, the harmonics of the inputs are detected and transmitted, which roughly indicates in which frequency the non-zeros are likely to be. This harmonics information is used for modifying the Rice parameters and also for zero run-length coding. This coder codes inputs for 320 samples per frame and foresees 320 samples per frame at a 16-kHz sampling rate. Thus 40 ms of algorithmic delay occurs.

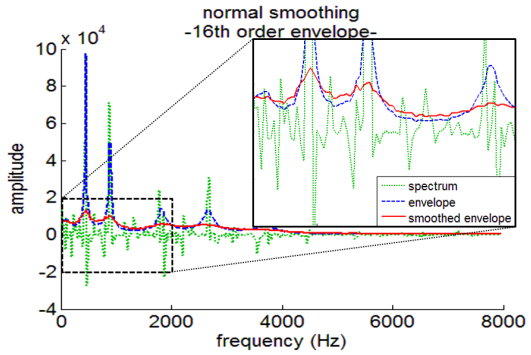
We compared the performance of the coder explained above using and without using the warped envelopes. Ten seconds of audio signals each from six categories in the RWC music database, a harpsichord piece for classic, a synthesizer piece for house, a piano piece for jazz, a guitar piece for pops, and male/female vocal pieces for vocal, were compressed into 16 kbps by the coder under the two conditions. We also compressed the inputs by AMR-WB+ for a benchmark. AMR-WB+ compresses the inputs for 1024 samples per frame at a 16-kHz sampling rate, and 72 ms of algorithmic delay occurs.

ITU-R BS.1534-1 MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) was carried out for the evaluation. Seven participants were presented the original signals for references, the signal compressed under each of the three conditions, and 3.5-kHz band-limited ones for anchors. The closeness of the signals to the references were graded from zero to a hundred points.

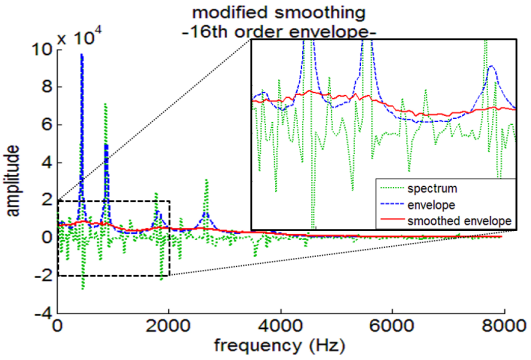
Fig. 7(a) displays the item-wise difference of scores in the MUSHRA test with the confidential intervals of 95 %. The coder described above showed almost the same performance as AMR-WB+ in scores. Fig. 7(b) shows the item-wise difference of the improvement in the scores by using the warping. Indeed, the influence of the warping varied by the items, but all the scores improved on average. Furthermore, in three out of the six items, qualities were enhanced at the significance level of 5 %. This results proves that the warping resulted in an enhancement of the qualities.

5. SUMMARY

We devised a representation of envelopes with warped resolution, which approximates frequency warping and its inverse by sparse and non-negative matrices. The resolution-warped envelopes showed almost the same fidelity as those of Mel-LPC method despite less computational complexity for extraction. In addition, we pointed out the necessity of modifying the perceptual weighting when applying it to the warped



(a) Smoothed by perceptual weighting eq. (2).



(b) Smoothed by the modified weighting eq. (13).

Fig. 5. The resolution-warped envelope smoothed by each method. The blue dashed line and the red solid line respectively indicate the envelope before and after smoothing. The LP order of the envelopes was 16.

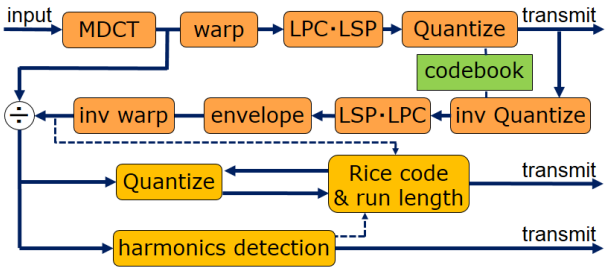


Fig. 6. Outline of the TCX based encoder.

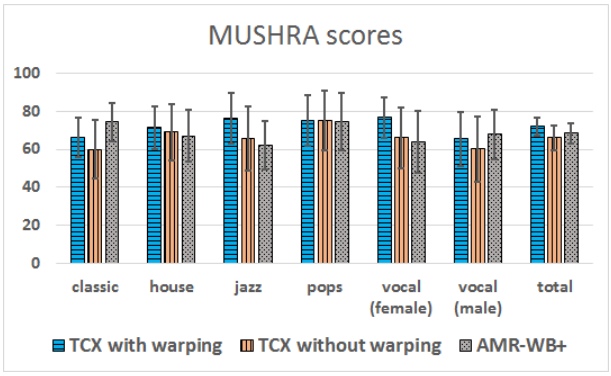
envelopes. Moreover, the enhancement in subjective quality by the warping was proven statistically.

A future issue is to verify under which conditions the warping has greater influence on enhancing the quality of the coding. We used the Mel-frequency scale for warping in this paper, but the scale for the warping is also still left for further modifications.

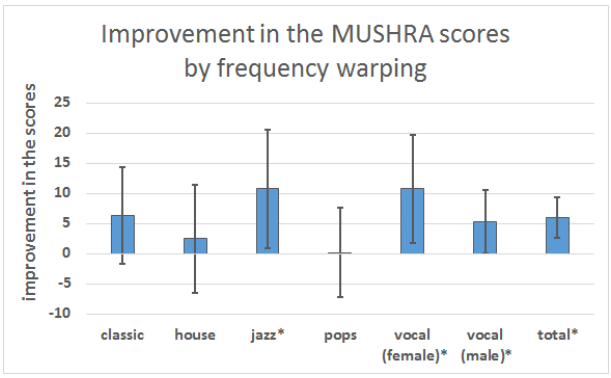
6. REFERENCES

[1] 3GPP TS 26.290 version 11.0.0 Release 11, 3GPP, 2012.

[2] M. Neuendorf, et al., "MPEG Unified Speech and Audio Coding - The ISO/MPEG standard for high-efficiency audio coding of all content types", AES, 2012.



(a) MUSHRA scores. The three types of bars represents, respectively from the left, TCX coder with warping, TCX coder without warping, and AMR-WB+.



(b) Improvements in the scores by using warping.

Fig. 7. Results of the subjective evaluation with confidence interval of 95 %. Items with an asterisk in the label had a difference at the significance level of 5 %.

[3] S. Wabnik, et.al., "Frequency warping in low delay audio coding," ICASSP, IEEE, vol.3, pp. /181- /184, 2005.

[4] H.W.Strube, "Linear prediction on a warped frequency scale," J.Acoust.Soc.Am., vol.68, no.4, pp.1071-1076, 1980.

[5] K. Tokuda, "Spectral estimation of speech by Mel-generalized cepstral analysis," Electron. Commun. Japan, vol.3, no.2, pp.30-43, 1993.

[6] K. Koishida, et al., "A wideband CELP speech coder at 16 kbit/s based on Mel-generalized cepstral analysis," ICASSP, IEEE, vol. 1, pp.161-164, 1998.

[7] G. Fuchs, et al., "MDCT-Based coder for highly adaptive speech and audio coding," EUSIPCO, IEEE, pp.1264-1268, 2009.

[8] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," Electron. Commun. Japan, vol. 53-A, pp. 36-43, 1970.

[9] H. Kameoka, et al., "Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelopes," IPSJ SIG Technical Reports, vol.2006-MUS-66, pp.77-84, Aug. 2006.

[10] http://sp-tk.sourceforge.net/ (as of Jan. '14).

[11] R. F. Rice, "Some practical universal noiseless coding techniques - part I-III," JPL Technical Report, vol. JPL-79-22, JPL-83-17, JPL-91-3, 1979, 1983, 1991.