A STUDY ON PERCEPTUAL DISTANCE MEASURE FOR PHASE SPECTRUM OF STIMULI

Hideki BANNO[1] *Kazuya TAKEDA*[1] *Fumitada ITAKURA*[2]

[1]Graduate School of Engineering, Nagoya University / CIAIR,
[2]Center for Information Media Studies, Nagoya University / CIAIR
Furo-cho 1, Chikusa-ku, Nagoya-shi, 464-8603, Japan

ABSTRACT

This paper describes a perceptual distance measure for phase spectrum based on results from a subjective experiment using stimuli. The stimuli have flat amplitude spectrum and, in a particular frequency band, have a certain group delay value. The experiment was performed using stimuli with different group delay peak values where the group delay center frequencies are fixed, and their associated group delay bandwidths are also fixed. It was found that when the peak values of stimuli are between -1 ms and 2 ms, they are perceived to be zero phase regardless of their center frequencies and bandwidths. Moreover, when the peak values are less than -8 ms or more than 10 ms and the bandwidths are less than 1 ERB, each of the stimuli are perceived to be similar. Based on these perceptual similarity results, we introduce an ellipsoidal function to estimate the similarity scores with a simple equation. It is found that the estimated similarity scores well approximate the subjective similarity scores.

1. INTRODUCTION

In most analysis-synthesis systems, zero or minimum phase reconstruction of speech is implemented [1] based on the assumption that, in human auditory perception phase information is not very important [2]. Honda and Moriya [3] have proposed a speech coder using phase equalization where low bit rate compression and high quality speech was obtained. However, in general the synthesized speech from such methods is reported to have a *buzziness* quality to it. Furthermore, most of these analysis-synthesis systems have a problem in making voiced/unvoiced decisions. Dealing with these problems has been one of the key issues in the speech synthesis field.

One of the most general methods in speech coding is to encode a residual of LPC called CELP [4]. The LPC residual includes phase information and amplitude information that can not be represented by LPC coefficients. Thus, the quality of the LPC residual-based synthesized speech is much better than that of the assumed zero phase or minimum phase synthesized speech. In addition, the residual based excitation has a merit that it does not require voicing information. Therefore, the LPC residual-based method is widely used in speech coding. However, the bit rate is inevitably quite more than that of the analysis-synthesis systems because of difficulties in dealing with the residual signal. Furthermore, it is also difficult to apply this model to speech modification, such as in speech transformation and speech morphing [5].

A speech manipulation method, proposed by Kawahara et al. [6] named STRAIGHT provided another evidence that phase information plays an important role in high quality speech synthesis. STRAIGHT is essentially a channel VOCODER using minimum phase impulse responses to resynthesize speech. However, each impulse response is modified using an all-pass filter which has random group delay characteristics in higher frequency region. By this randomization of the group delay, naturalness of STRAIGHT is much improved without losing flexibility in the vocoder. Even though STRAIGHT provides high quality synthesized speech, it is still possible to discriminate the difference between the original speech and the synthesized speech. STRAIGHT also has the voiced/unvoiced decision problem as the other analysis-synthesis methods do.

In a study of auditory perception, several researchers reported that the human auditory system is not deaf to the phase difference [7, 8, 9, 10]. Patterson also reported that and proposed a model that simulates human auditory perception properly [11, 12]. However, it is not easy to apply this model directly to speech generation.

Previously, we have proposed a short-time phase representation for application to speech coding or speech transformation [13]. This research showed that the proposed method can effectively reduce the phase parameter in such applications. However, for further reduction of the phase representation, quantitative analysis on the characteristics in phase difference perception is indispensable. In this paper, we propose a parametric representation of the phase spectrum based on a type of group delay representation and measure how phase difference is perceived by human subjects. We also propose a computational method to estimate perceptual phase distance measures between a stimuli pair.

2. STIMULI DESIGN

In this research, the phase spectrum of stimuli is designed in the group delay domain, which allows the control of the closeness to zero-phase model in the stimuli. All stimuli have a flat amplitude spectrum. The phase spectrum is computed from the group delay spectrum given by the following single peak function,

$$D(f) = \tau_0 \exp(-\pi((f - f_c)/f_b)^2), \tag{1}$$

where, f_c is the center frequency, f_b is the parameter that defines bandwidth, and τ_0 is the group delay value at the center frequency. If τ_0 is equal to 0 at all frequencies, the stimulus becomes a zero phase (impulse) signal. The bandwidth parameter f_b is expressed as a function of ERB (Equivalent Rectangular Bandwidth) which indicates bandwidth of the auditory filter,

$$f_b = c_e ERB, \tag{2}$$

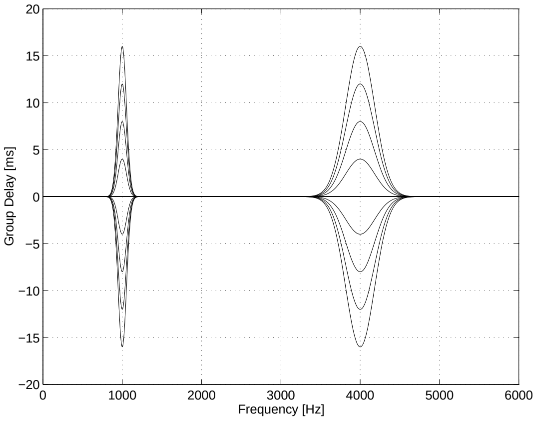

Fig. 1. Group delay spectra of the stimuli when the group delay value τ_0 is variable, the center frequency f_c is 1000 Hz or 4000 Hz, and the bandwidth parameter c_e is 1.0.

where, c_e is a set parameter. The ERB is computed as a function of the center frequency by

$$ERB = 24.7(4.37 f_c/1000 + 1). \quad (3)$$

Figure 1 shows group delay spectra of the stimuli when the group delay value τ_0 is variable, the center frequency f_c is 1000 Hz and the bandwidth parameter c_e is 1.0.

The stimuli design is performed by the method described in [14]. This method realizes causal group delay filters by cascading second-order all-pass filters. The delay component corresponding to the filter order is added to the ideal group delay through this procedure. Note that the realized group delay is compensated to remove the delay component associated with all realized group delays in this paper. Since the group delay is accurate within less than about 0.1 ms, we ignore the error found in these filters.

The generated group delay filter is used for a single period of a stimulus. A stimulus signal is obtained by overlap-adding this single period signal. The fundamental frequency of the overlap-addition is chosen to be 50 Hz. This quite low fundamental frequency is to enable easier perception of phase differences of the stimuli, as described in some reports [7, 9]. The stimulus has 5 ms linear taper at the rise and the fall of the waveform.

3. EXPERIMENT: PERCEPTION OF PEAK GROUP DELAY

Using the above representation of stimuli by group delay, we have investigated the potential of human phase perception. In this experiment, both f_c and c_e were set a priori, and the group delay peak value τ_0 was variable.

3.1. Conditions

Pairs of the created stimuli were presented to listeners in random order. Listeners responded to the similarity of the stimulus pair as one of five grades shown in Table 1. Four listeners with normal hearing participated in this experiment. The tests were performed under the conditions listed in Table 2. Headphone spectral characteristics were compensated for by using the inverse filter in advance. The volume of the headphone is adjusted so that the maximum sound pressure level at the entrance of an ear canal is equal to 76 dBA. The interval between the stimuli in each trial was 1 second.

Table 1. The evaluation measure on the hearing test.

5	I cannot discriminate between the stimuli at all.
4	The stimuli are very similar.
3	The stimuli are quite similar.
2	The stimuli are not similar.
1	The stimuli are completely different.

Table 2. Conditions of creating stimuli in Experiment.

Sampling frequency	16000 Hz
Fundamental frequency	50 Hz (20 ms)
Duration	3 sec
Center frequency (f_c)	1000 Hz, 4000 Hz
Bandwidth parameter (c_e)	0.6, 2.0
Peak value of group delay (τ_0)	-16 ms to 16 ms

3.2. Results and Discussion

Figure 2 shows the relationship between the group delay peak values and the similarity measure, when the group delay peak value of a stimulus is close to zero, that is, the stimulus is close to zero phase. The center frequency of this figure is 1000 Hz. Although there is no figure where the center frequency is 4000 Hz, the figure of that case shows almost same tendencies as the 1000 Hz case. In this figure, each plot corresponds to the similarity between the stimulus with a specific peak value and the stimulus with a peak value indicated in the horizontal axis. The similarity means the subjective similarity of the stimuli in MOS.

From this figure, it can be concluded that stimuli are perceived to be similar regardless of f_c and c_e, when the stimulus has small delay (-1 ms $< \tau_0 <$ 2 ms). Furthermore, it is deduced that the difference between zero phase stimulus and the stimulus with large delay ($\tau_0 < -8$ ms, 10 ms $< \tau_0$) is easily perceived regardless of f_c and c_e. These results indicate that the human auditory system is quite sensitive to the difference between zero phase and non-zero phase stimuli.

Figure 3 shows the relationship between the group delay peak values and the similarity, when the group delay peak value of a stimulus is far from zero, that is, the stimulus is non-zero phase.

From this figure, the difference between the stimulus with large delay ($\tau_0 < -8$ ms, 10 ms $< \tau_0$) and the stimulus with small delay (-1 ms $< \tau_0, \tau_0 <$ 2 ms) can easily be perceptually discriminated. Thus, it also can be concluded that the auditory system is quite sensitive to the difference between zero phase and non-zero phase.

Moreover, when the bandwidth parameters c_e is 0.6, all the stimuli with long delay are perceived to be similar. This means that the auditory system is less sensitive in phase discrimination of non-zero phase components with narrow group delay bandwidths. On the other hand, the case where c_e is 2.0 shows a lower similarity than the case where c_e is 0.6.

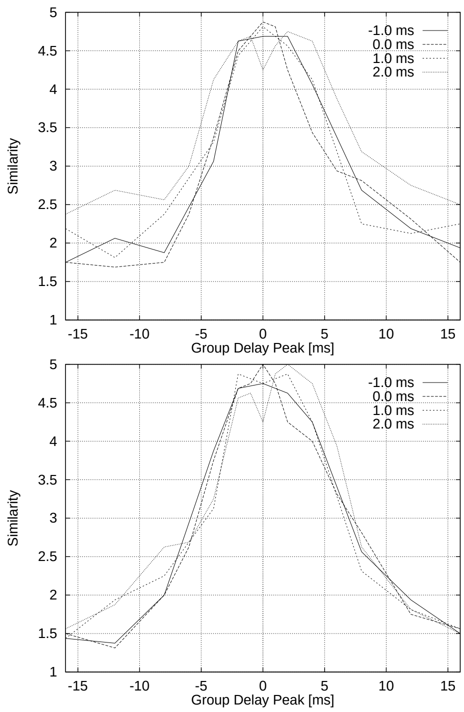

Fig. 2. Relationship between the group delay peak value τ_0 and the similarity measure where $c_e = 0.6$ (top) and $c_e = 2.0$ (bottom). In both panels, τ_0 is close to 0 ms and $f_c = 1000$ Hz.

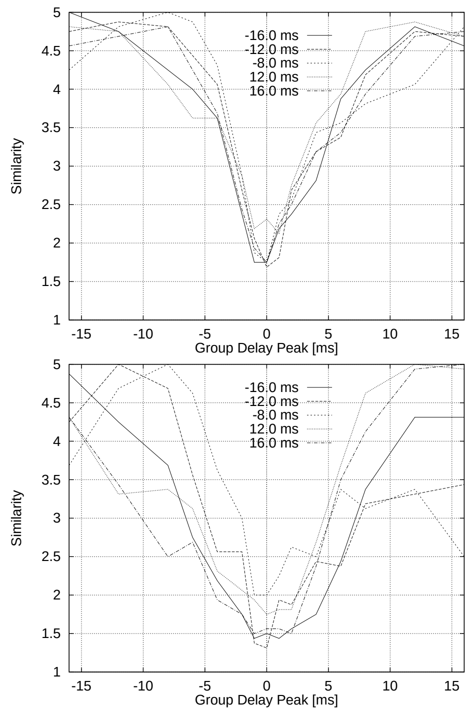

Fig. 3. Relationship between the group delay peak value τ_0 and the similarity measure where $c_e = 0.6$ (top) and $c_e = 2.0$ (bottom). In both panels, τ_0 is far from 0 ms and $f_c = 1000$ Hz.

4. PERCEPTUAL DISTANCE MEASURE OF PHASE SPECTRUM BASED ON AN ELLIPSOIDAL MAPPING

The above experiment showed that when the group delay peak values of stimuli are between -1 ms and 2 ms, they are perceived to be zero phase regardless of their center frequencies and bandwidths. Is also showed that the zero phase stimuli is not similar to the non-zero phase stimuli. Moreover, when the stimuli is non-zero phase and the bandwidths are 0.6 ERBs, each stimuli pair is perceived to be similar.

Based on these perceptual similarity results, we introduce an ellipsoidal function to explain such findings with a simple equation. We refer to this ellipsoidal function as the *similarity ellipse* shown in Figure 4. Stimuli are assigned to the ellipse boundary at uneven intervals according to the peak values of the stimuli. In this ellipse, zero phase stimuli are assigned to around zero radian, and non-zero phase stimuli are assigned to around π radian. Stimuli with positive group delay and negative group delay are assigned to upper half area and lower half area, respectively. The similarity can be estimated by the Euclidean closeness between locations on the ellipse boundary. Limiting the perceptual similarity measures of the stimuli to an ellipsoidal boundary formulates the framework for our study.

A similarity score in our experiments between a stimuli pair can be estimated by

$$s(\tau_1, \tau_2) = 5 - d(\tau_1, \tau_2), \quad 1 \leq s(\tau_1, \tau_2) \leq 5 \qquad (4)$$

where, τ_1 and τ_2 are peak values of two stimuli, respectively and $d(\tau_1, \tau_2)$ is a distance between the stimuli given by

$$d(\tau_1, \tau_2) = |\boldsymbol{x}(\tau_1) - \boldsymbol{x}(\tau_2)|, \quad 0 \leq d(\tau_1, \tau_2) \leq 4. \qquad (5)$$

Here, $\boldsymbol{x}(\tau)$ is a vector on an ellipse boundary as a function of a group delay value, described by

$$\boldsymbol{x}(\tau) = (a \cos \phi(\tau), b \sin \phi(\tau)), \quad a, b > 0 \qquad (6)$$

where, a and b are parameters defining the shape of the ellipse. The function $\phi(\tau)$ is a nonlinear function mapping a change in group delay varying between $-\pi$ and π, where τ is limited between -16ms to 16ms. This function is based on the sigmoidal function given by

$$\phi(\tau) = \frac{2\pi}{1 + \exp(-(\tau - u)/v)} - \pi, \quad u, v > 0 \qquad (7)$$

where, u describes the horizontal shift and v the slope of the function.

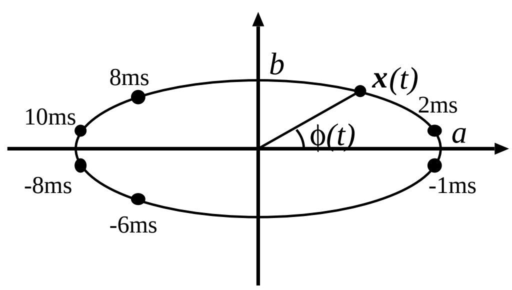

Fig. 4. The similarity ellipse for the estimation of experimental similarity scores.

Table 3. Optimum parameter values and errors.

Stimuli type	a	b	u	v	Error
(1)	1.5	0.4	0.4	4.2	0.31
(2)	1.9	0.8	0.3	6.2	0.30
(3)	1.8	0.4	1.0	5.8	0.27
(4)	2.2	0.7	0.8	6.8	0.31

The parameters a, b, u and v were optimized based on the minimum least square error criterion for $s(\tau_1, \tau_2)$ to fit with the similarity scores obtained by the subjective experiment. The optimization was performed in the following cases independently, (1) $f_c = 1000$ Hz and $c_e = 0.8$, (2) $f_c = 1000$ Hz and $c_e = 2.0$, (3) $f_c = 4000$ Hz and $c_e = 0.8$, and (4) $f_c = 4000$ Hz and $c_e = 2.0$. Table 3 shows the optimum parameters and errors between the estimated similarity scores and the similarity scores obtained by subjective experiments. The error obtained is about 0.3 in similarity score as shown in this table. This error is quite small considering that the maximum similarity score is five and the subjective experiment has an inherent variance associated with them.

Figure 5 shows the resultant similarity score representation by this optimization method and the similarity scores obtained by the subjective experiment. It is found that the estimated similarity scores well approximate the similarity scores obtained by the subjective experiment. These results imply that the human auditory system perceives the phase spectral differences based on a simple ellipsoidal function.

5. CONCLUSIONS

A subjective experiment for investigating the characteristic of human phase perception was performed using stimuli that have flat amplitude spectra and various group delay values only in a particular frequency band. It was clarified that the human auditory system is sensitive to the difference between zero and non-zero phase. It was also clarified that the human auditory system is insensitive to the the difference among the non-zero phase stimuli when the group delay bandwidth of the stimulus is 0.6 ERB. To explain these findings with a simple equation, we introduced an ellipsoidal function. The similarity scores estimated by using the function well approximate the similarity scores obtained by subjective experiments. Future work includes the establishment of a perceptual distance measure for phase spectrum which is applicable to speech processing.

Acknowledgments This work has been partially supported by Grant-in-Aid for COE Research (No. 11CE2005) and Core Research for Evolutional Science and Technology (CREST) program of Japan Science and Technology Corporation (JST).

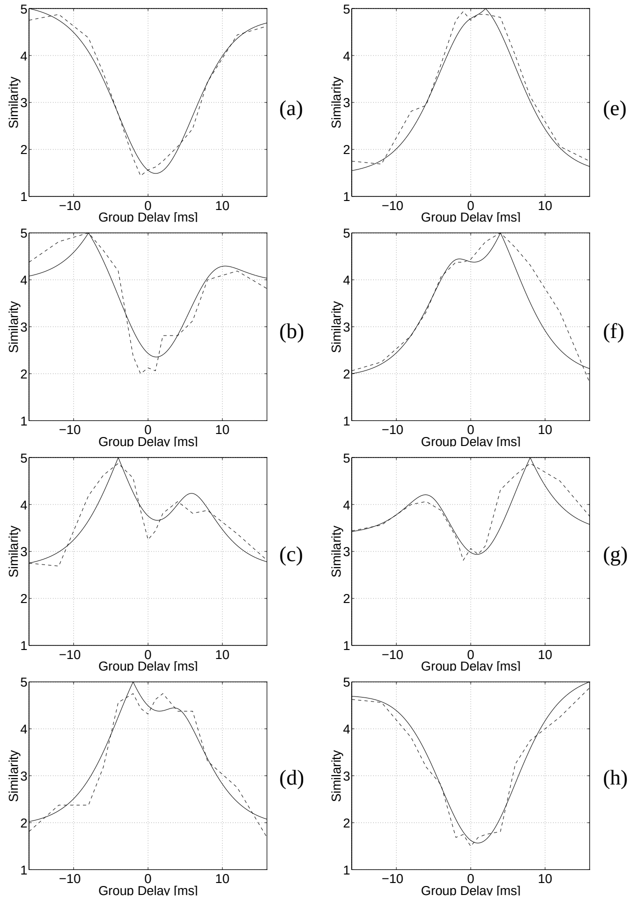

Fig. 5. Relationship between the group delay peak value τ_0 and the similarity score when $f_c = 4000$ Hz and $c_e = 0.6$. The solid line and the dashed line describe the estimated similarity score and the similarity score obtained by the subjective experiment, respectively. Each panel shows the case where the group delay peak value τ_0 is (a) -16 ms, (b)-8 ms, (c)-4 ms, (d)-2 ms, (e)2 ms, (f)4 ms, (g)8 ms and (h)16 ms.

6. REFERENCES

[1] J.L.Flanagan, "Speech Analysis, Synthesis, and Perception," (*Sprinnger-Verlag*, New York, 1972).

[2] H.Fletcher, "Loudness, Pitch and the Timbre of Musical Tones and Their Relation to the Intensity, the Frequency and the Overtone Structure," *J. Acoust. Soc. Amer.* **6** (2), 59-69 (1934).

[3] T.Moriya and M.Honda, "Speech Coder Using Phase Equalization and Vector Quantization," *Proc. ICASSP 1986*, 1701–1704 (1986).

[4] M.R.Schroeder, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," *Proc. ICASSP 1985*, 937–940 (1985).

[5] H.Banno, S.Kajita, K.Takeda, F.Itakura and K.Shikano, "Speech Morphing by Progressive Interpolation of Spectra," *Proc. ICSP 1997*, 149–153 (1997).

[6] H.Kawahara, "Speech Representation and Transformation using Adaptive Interpolation of Weighted Spectrum: VOCODER Revisited," *Proc. ICASSP 1997*, 1303–1306 (1997).

[7] J.C.R.Licklider, "Effects of Changes in the Phase Pattern upon the Sound of a 16-Harmonic Tone," *J. Acoust. Soc. Amer.* **29**, p.780 (1957).

[8] M.R.Schroeder, "New Results Concerning Monaural Phase Sensitivity," *J. Acoust. Soc. Amer.* **31**, p.1579 (1959).

[9] R.Plomp and H.J.M.Steeneken, "Effect of Phase on the Timbre of Complex Tones," *J. Acoust. Soc. Amer.* **46** (2), 409–421 (1969).

[10] J.Blauert and P.Laws, "Group delay distortion in electroacoustical systems," *J. Acoust. Soc. Amer.* **63** (5), 1478–1483 (1978).

[11] R.D.Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Amer.* **82** (5), 1560–1586 (1987).

[12] R.D.Patterson, "The sound of a sinusoid: Time-interval models," *J. Acoust. Soc. Amer.* **96** (3), 1419–1428 (1994).

[13] H.Banno, J.Lu, S.Nakamura, K.Shikano and H.Kawahara, "Efficient Representation of Short-time Phase Based on Group Delay," *Proc. ICASSP 1998*, 861–864 (1998).

[14] P.A.Bernhardt, "Simplified Design of High-Order Recursive Group-Delay Filters," *IEEE Trans. ASSP* **28**, 498–503 (1980).