Identifying Appropriate Dynamic Time Warping, *DTW*, Metrics to Evaluate the Quality of Vocoder Time-Stretched Speech Phrases

Mai Tanaka, and Michael R. Smith, *Senior Member, IEEE*

***Abstract*— To improve comprehension in a noisy environment, we suggest sending short speech phrases slowed by a phase vocoder directly to seniors' hearing aids. Dynamic time warping, DTW, metrics for evaluating the speech quality of vocoder modified phrases are compared.**

I. Introduction

Many factors deteriorate the intelligibility and quality of speech, resulting in unsuccessful and frustrating conversations. There are compensatory actions that people can take to make speech easier to understand. For example, people talking in a noisy restaurant speak at a slower rate [1].

The phase vocoder is designed to use various time and frequency domain manipulations to slow down, time stretch, an audio signal with minimal pitch distortions [2]. We suggest that using on-demand, phase vocoder manipulation of short speech phrases sent directly to hearing aids might have a role in increasing the level of comprehension for seniors, and reduce their social isolation.

Measures of the speech intelligibility and quality of time-stretched signals could be made using subjective ratings by human subjects [3]. However, learning effects would bias repeated attempts to measure the intelligibility of different stretched variants of the same audio. We propose evaluating vocoder performance using an automated audio observer based on dynamic time warping, *DTW*.

II. Method

DTW is a pattern matching algorithm designed to identify similarities between signals that are locally time distorted. As this application involves globally time distorted signals, calibration of the effectiveness of *DTW* metrics was made using a known simulated single-tone phrase that could be exactly time stretched (Fig 1) with no pitch distortion. Two *DTW* metrics were evaluated.

- Direct *DTW* distances obtained by comparing the original and time stretched signals, and
- *MFCC-DTW* which uses *DTW* distances calculated between the spectral representations of the original and time stretched speech using Mel Frequency Cepstral Coefficients (*MFCC*) [3]. The envelope of the single tone signal in Fig. 1 was used as an ideal representation of the tones *MFCC*

*Supported by Natural Sciences and Engineering Research Council of Canada and the Campbell McLaurin Foundation for Hearing Research.

M. Tanaka is with the Department of Electrical and Computer Engineering, University of Calgary, Canada (e-mail: tanakam@ucalgary.ca).

M. R. Smith is with the Department of Electrical and Computer Engineering and the Department of Radiology (e-mail: smithmr@ucalgary.ca; phone: +1-403-220-6142).

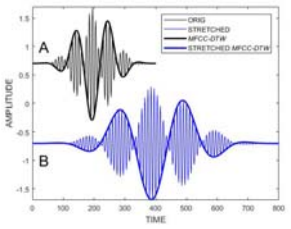

Fig. 1. Signal and envelope coefficients for A) original and B) ideal time stretched signals.

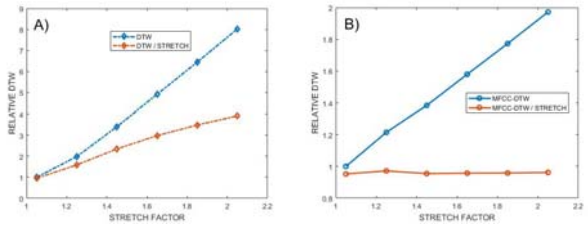

Fig. 2) A) Standard *DTW* and B) *Envelope-DTW* metrics for ideal vocoder time-stretched phrases.

III. Results and Conclusion

An appropriate metric reflecting vocoder distortion would show no changes for ideally time stretched signals. Neither *DTW* nor *Envelope-DTW* meet this criterion, as seen in the upper blue curves in Figs 2A and 2B. We propose a new *normEnvelope-DTW* metric which adjusts for the level of imposed time stretching. The lower curve of Fig. 2B, red line, indicates this new metric has the required behaviour .

Future work will extend these metrics to include *normMFCC-DTW*. A band-by-band evaluation will be made of vocoder distortions introduced into voice phrases from the NOIZEUS data base [4]. *MFCC* coefficients represent frequency bins with bandwidths similar to the hearing behavior of the human ear [3]. This is important given that high frequency hearing loss has a significant impact on speech comprehension. We suggest that potential frequency specific distortions introduced by the phase vocoder would be well differentiated using the *normMFCC-DTW* metric.

References

[1] S. Liu and F.-G. Zeng, "Temporal properties in clear speech perception," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 424–432, 2006.

[2] A. De Götzen, N. Bernardini, and D. Arfib, "Traditional implementations of a phase-vocoder: the tricks of the trade," in *COST-G6 Conference on Digital Audio Effects*, 2000, pp. 1–7.

[3] I. Kraljevski, S. Chungurski, et al., "Perceived speech quality estimation using DTW algorithm," *Telfor J.*, vol. 1, no. 1, pp. 22–25, 2009.

[4] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7–8, pp. 588–601, 2007.