ELSEVIER

Speech Communication 36 (2002) 343–374

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A comparison of spectral smoothing methods for segment concatenation based speech synthesis ☆

David T. Chappell [b], John H.L. Hansen [a,b,*]

[a] *Robust Speech Processing Laboratory (RSPL), Center for Spoken Language Research (CSLR), Room E265, University of Colorado, 3215 Marine St., P.O. Box 594, Boulder, CO 80309-0594, USA*

[b] *Department of Electrical Engineering, P.O. Box 90291, Duke University, Durham, NC 27708-0291, USA*

Received 21 April 1999; received in revised form 24 May 2000; accepted 15 December 2000

Abstract

There are many scenarios in both speech synthesis and coding in which adjacent time-frames of speech are spectrally discontinuous. This paper addresses the topic of improving concatenative speech synthesis with a limited database by proposing methods to smooth, adjust, or interpolate the spectral transitions between speech segments. The objective is to produce natural-sounding speech via segment concatenation when formants and other spectral features do not align properly. We consider several methods for adjusting the spectra at the boundaries between waveform segments. Techniques examined include optimal coupling, waveform interpolation (WI), linear predictive parameter interpolation, and psychoacoustic closure. Several of these algorithms have been previously developed for either coding or synthesis, while others are enhanced. We also consider the connection between speech science and articulation in determining the type of smoothing appropriate for given phoneme–phoneme transitions. Moreover, this work incorporates the use of a recently-proposed auditory-neural based distance measure (ANBM), which employs a computational model of the auditory system to assess perceived spectral discontinuities. We demonstrate how actual ANBM scores can be used to help determine the need for smoothing. In addition, formal evaluation of four smoothing methods, using the ANBM and extensive listener tests, reveals that smoothing can distinctly improve the quality of speech but when applied inappropriately can also degrade the quality. It is shown that after proper spectral smoothing, or spectral interpolation, the final synthesized speech sounds more natural and has a more continuous spectral structure.

Keywords: Speech synthesis; Speech coding; Spectral smoothing; Spectral interpolation

1. Introduction

When speech is produced naturally by a human, there is a measurable degree of continuity between phone segments. This degree of continuity is related to the physical movement and placement of the vocal system articulators. When speech is produced artificially, such as in segment-based synthesis or in low-bit-rate coding, the same phone-to-phone continuity may not exist.

Speech synthesis, coding, and voice transformation can benefit from improvements in spectral smoothing. There are a number of scenarios in which the spectral structure of speech at adjacent

☆ This work was supported in part by SPAWAR under grant No. N66001-92-0092.

* Corresponding author. Tel.: +1-303-735-5148; fax: +1-303-735-5072.

E-mail address: John.Hansen@colorado.edu (J.H.L. Hansen).

0167-6393/02/$ - see front matter
PII: S0167-6393(01)00008-5