Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 76 (2016) 93–111

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Formant measurement in children's speech based on spectral filtering

Brad H. Story*, Kate Bunton

Speech Acoustics Laboratory, Department of Speech, Language, and Hearing Sciences, University of Arizona, P.O. Box 210071, Tucson, AZ 85721, United States

Received 17 July 2015; received in revised form 7 November 2015; accepted 13 November 2015
Available online 26 November 2015

Abstract

Children's speech presents a challenging problem for formant frequency measurement. In part, this is because high fundamental frequencies, typical of a children's speech production, generate widely spaced harmonic components that may undersample the spectral shape of the vocal tract transfer function. In addition, there is often a weakening of upper harmonic energy and a noise component due to glottal turbulence. The purpose of this study was to develop a formant measurement technique based on cepstral analysis that does not require modification of the cepstrum itself or transformation back to the spectral domain. Instead, a narrow-band spectrum is low-pass filtered with a cutoff point (i.e., cutoff "quefrency" in the terminology of cepstral analysis) to preserve only the spectral envelope. To test the method, speech representative of a 2–3 year-old child was simulated with an airway modulation model of speech production. The model, which includes physiologically-scaled vocal folds and vocal tract, generates sound output analogous to a microphone signal. The vocal tract resonance frequencies can be calculated independently of the output signal and thus provide test cases that allow for assessing the accuracy of the formant tracking algorithm. When applied to the simulated child-like speech, the spectral filtering approach was shown to provide a clear spectrographic representation of formant change over the time course of the signal, and facilitates tracking formant frequencies for further analysis.

Keywords: Formant; Vocal tract; Speech analysis; Children's speech; Speech modeling.

1. Introduction

The formant frequencies present in a speech signal are regions of spectral prominence for which acoustic energy has been enhanced, and provide cues for the perception of both vowels and consonants. Formants are the resultant effect of the interaction of vocal tract resonances with a source of sound, and thus provide an "acoustic window" to the shape of the vocal tract, albeit indirect. For adult speech, wide-band spectrography and linear prediction (LP) techniques (e.g., Makhoul, 1975; Markel and Gray, 1976) can provide reasonably accurate measurements of the formants (Monsen and Engebretson, 1983; Vallabha and Tuller, 2002). In part, this is because the ample harmonics produced by the voice source adequately sample the vocal tract transfer function and express it clearly as the envelope of the speech spectrum. On the other hand, children's speech is typically characterized by high fundamental frequencies (e.g., 250–600 Hz) that generate widely spaced harmonic components, producing an apparent undersampling of the vocal tract transfer function (c.f., Kent, 1976; Lindblom, 1962). It then becomes difficult to measure the formant frequencies from the spectral envelope because the envelope peaks are strongly influenced by individual harmonic amplitudes rather than by a collective effect of many closely-spaced harmonics. In addition, children may produce a breathy voice quality characterized by low amplitude upper harmonics and a significant noise component due to glottal turbulence (Ferrand, 2000; Glaze et al., 1988).

The problem of estimating formants from speech with high fundamental frequencies (f_o) is well known (cf. Fant, 1968; Kent, 1976; Lindblom, 1962; 1972) and has been addressed with a variety of techniques and methods. With wide-band spectrography, for example, the effective filter bandwidth can be increased to obscure individual harmonic components such that broad formant peaks can be observed (Eguchi and Hirsh, 1969; Bennett, 1981; Kent, 1976; Lindblom, 1972) , but potentially with considerable error as pointed out by Lindblom (1972). Modifications to linear prediction methods have also been proposed to deal

* Corresponding author. Tel.: +1 5206269528; fax: +1 5206219901.
E-mail address: bstory@email.arizona.edu, bstory@u.arizona.edu (B.H. Story).

http://dx.doi.org/10.1016/j.specom.2015.11.001
0167-6393/

with formant measurement in high-f_o speech (cf., El-Jaroudi and Makhoul, 1991; Hermansky et al., 1984; Ma et al., 1993). More recently, (Alku et al., 2013) proposed a weighted linear prediction technique in which the main points of excitation within each glottal cycle are attenuated. This has the effect of giving more weight to the portions of each cycle that contain information about vocal tract resonances rather than the voice source, and results in better estimates of formant frequencies. Liu and Shimamura (2015) reported a similar technique but without the need to identify glottal closure epochs.

Undersampling the vocal tract transfer function in high-f_o speech can be mitigated to some degree by varying the fundamental frequency over the time course of an utterance. This has the effect of sweeping the f_o and associated harmonic components through the resonance peaks in the transfer function, thus producing a more complete excitation of the formant structure, albeit over an adequately long temporal window. White (1999) reported a formant measurement technique in which 11 year-old children were asked to produce a vowel, either spoken or sung, while simultaneously shifting their f_o from low to high frequency. The duration of the recordings was 1–2 s and formants were identified from a narrow-band spectrogram as the points at which the harmonic amplitudes were highest; these coincided with the points in time at which a particular harmonic passed through a resonance peak in the vocal tract transfer function. This is perhaps a useful method, but relies on the ability of the talker to perform the unusual task of maintaining a static vocal tract configuration during a pitch glide, and does not lend itself to analysis of time-varying speech. Wang and Quatieri (2010) similarly exploited f_o changes to develop a signal processing technique for detecting the vocal tract resonances in high-f_o speech, but relied on the natural variation of f_o in human speech rather than deliberately asking talkers to produce f_o glides. Using localized 2D Fourier transforms of the temporal-spatial variation of speech, they showed an improved separation of the voice source and vocal tract filter when the f_o was changing.

Cepstral analysis is an alternative approach to measuring formants in high-f_o speech. The envelope of the log spectrum of a speech segment can be considered analogous to a low frequency modulation of a waveform, whereas the individual harmonics or noise components can be regarded as an analogy to a carrier signal. Thus, calculation of the log spectrum of the *initial log spectrum* results in yet another kind of spectrum, called the *cepstrum* (Bogert et al., 1963), that separates the envelope from the harmonics and higher frequency noise. The cepstrum can be modified such that only the portion related to the envelope is retained, and then transformed back to the spectral domain. The result is an estimate of the spectral envelope, the peaks of which are representative of the formants (cf. Childers et al., 1977). As has been shown (Fort and Manfredi, 1998), cepstral filtering can be enhanced by allowing the filter (or "lifter") length to be dependent on the fundamental frequency within a given time frame, and using a chirp Z-transform to improve the resolution for finding spectral peaks. Rahman and Shimamura (2005) have also improved formant tracking in high-f_o signals by applying linear prediction to the portion of the cepstrum related to the vocal tract impulse response.

The purpose of this study was to develop and test a technique for visualizing and measuring formants in children's speech with a wide range of variation of f_o, vocal tract resonances, and distribution of harmonic and noise-like energy. The method is conceptually based on cepstral analysis (Bogert et al., 1963), but does not require modification of the cepstrum itself or transformation back to the spectral domain. Instead, the narrow-band spectrum over any given time-window is low-pass filtered with a selected cutoff point (i.e., cutoff "quefrency" in the terminology of cepstral analysis), determined by the time-dependent f_o, to preserve only the spectral envelope. Formants are then measured by applying a peaking-picking algorithm to the spectral envelope.

Assessing the accuracy of formant tracking algorithms applied to natural (recorded) speech can be problematic because the "true" answer is typically unavailable. That is, an algorithm will deliver measurements of the formant frequencies but whether they are reasonable estimates of the vocal tract resonances produced by the talker is unknown. A typical paradigm for testing is to apply the algorithm to synthetic speech for which the resonance frequencies are known *a priori* and compare them to the formant values determined by the algorithm. As illustrated in Fig. 1, a similar paradigm was used here by applying the spectral filtering method to artificial speech samples, representative of a 2–3 year-old child talker, that were produced with a computational model such that resonance frequencies could be calculated independently of the spectral filtering algorithm. The model allows for time-dependent variations in vocal tract shape, f_o, and degree of vocal fold adduction. In addition, the glottal flow signal is produced interactively with the propagating acoustic pressures in the vocal tract, and contains noise that emulates the effects of glottal turbulence. Thus, the generated audio samples include characteristics similar to those observed in children's speech and provide reasonably challenging cases for testing the algorithm (or any other algorithm designed to track formants).

The specific aims of the paper are to: (1) describe the model used to simulate child-like speech samples, (2) describe the spectral filtering algorithm, and (3) apply the algorithm to the simulated speech samples and compare results to the known values of vocal tract resonances. In addition, the spectrographic representation provided by the spectral filtering algorithm will be compared to that given by a conventional linear prediction algorithm.

2. Simulation of child-like speech

The speech production model depicted in Fig. 1 consists of two main components: (1) a kinematic representation of the medial surfaces of the vocal folds, and (2) a vocal tract airway defined by an area function. The vocal fold medial surfaces can be controlled to modulate the glottal airspace on a slow time scale for adduction and abduction maneuvers, as well as on a more rapid time scale to emulate vocal fold vibration at speech-like fundamental frequencies (Titze, 2006). When supplied with a subglottal pressure the modulation of the glottis produces a glottal airflow signal, $u_g(t)$, that provides the acoustic