Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 298–313

www.elsevier.com/locate/csl

Improvements to harmonic model for extracting better speech features in clinical applications☆

Meysam Asgari*, Izhak Shafran

Center for Spoken Language Understanding, Oregon Health & Science University, Portland, Oregon, United States

Received 14 April 2016; received in revised form 10 July 2017; accepted 31 August 2017
Available online 8 September 2017

Abstract

Acoustic properties of speech samples can provide important cues in the assessment of voice pathology and cognitive function. The goal of this study is to develop novel algorithms for robust and accurate estimation of speech features and employ them to build probabilistic speech models for characterizing and analyzing clinical speech. Toward this goal, we adopt a harmonic model (HM) of speech. We overcome certain drawbacks of this model and introduce an improved version of HM that leads us to accurate and reliable estimation of voiced segments, fundamental frequency, HNR, jitter, and shimmer. We evaluate the performance of our improved HM in the context of voicing detection and pitch estimation with other state-of-the-art techniques on the Keele data set. Through extensive experiments on several noisy conditions, we demonstrate that the proposed improvements provide substantial gains over other popular methods under different noise levels and environments. Next, we investigate the utility of developed measures on the speech-based assessment of cognitive impairments including clinical depression and autism spectrum disorder (ASD). Our preliminary results on two clinical tasks demonstrate the promise of our improved HM features in practical applications.

Keywords: Pitch tracking; Voice activity detection; Modified harmonic model

1. Introduction

Analysis of acoustic signals of the human voice has many purposes. Our voice reveals considerable insight into the structure and function of certain organs involved in speech and language production. For instance, sometimes the first symptom of a neurological disorder such as Parkinson's disease (PD) is a speech impairment (Duffy, 2000). PD can affect all components of speech production including breathing, laryngeal function, and articulation, as well as their coordination for the production of smooth speech. The resulting dysarthric speech often exhibits monotonous pitch, variable speech rate, and a harsh and breathy voice (Darley et al., 1969). Likewise, researchers have shown the effects of psychological disorders, such as depression, in patients' voices (Low et al., 2011). Speech pathologists have characterized depressed speech as monotonous, mono-loud, and low in range of pitch

☆ This paper has been recommended for acceptance by R. K. Moore.

* Corresponding author.
E-mail addresses: meysam_asgari@yahoo.com, asgari@ohsu.edu (M. Asgari).

http://dx.doi.org/10.1016/j.csl.2017.08.005
0885-2308/

frequency (Moore et al., 2004). Also, a number of studies have shown that emotional arousal considerably influences phonatory and articulatory aspects of the speech production system (Moses, 1954). In addition to PD and depression, autism spectrum disorder (ASD) is another example of a disorder that affects patients' voices. Children with autism often exhibit unusual pitch and intonation: for example, monotonous pitch, reduced stress, odd rhythm, large pitch range (Hubbard and Trauner, 2007), and even differences in the harmonic structure of their speech (Bonneh et al., 2010).

The severity of the aforementioned diseases is typically assessed subjectively by an expert practitioner and often requires the patient's presence at the clinic. This assessment is often costly and time-consuming, and can be burdensome in some situations, such as when a patient must undergo frequent reassessments. For several decades now, symptoms observed in the speech of patients with such diseases have motivated researchers to explore alternative approaches based on speech processing techniques. Researchers have measured these symptoms more objectively with the hope of augmenting or simplifying the assessment. It is often cheaper and easier to automatically elicit, record, and analyze speech than to conduct an in-person clinical assessment. Furthermore, speech-based assessment can be remotely administered and can be used to objectively monitor changes over time. Easier methods of assessment, such as automated screening and telemonitoring, can play a crucial role in the early detection of the aforementioned diseases. The main focus of this paper is developing novel algorithms for robust and accurate estimation of pitch-related features, and employing them to build probabilistic speech models for characterizing and analyzing clinical speech. There are a number of approaches in the time and frequency domains to estimate pitch-related features. Time domain methods often ignore frequency and amplitude variations of speech over the analysis frame. On the other hand, the resolution of the short time Fourier transform does not provide the necessary time-frequency resolution to capture the small amount of perturbation observed in, for example, Parkinson's disease (PD). As an alternative we adopt the harmonic model of speech, which has recently gained considerable attention. This model takes into account the underlying harmonic nature of voiced speech and decomposes it into a harmonic and a non-harmonic component. We overcome certain drawbacks of this model and introduce an improved version of HM that leads us to accurate and reliable estimation of voiced segments, fundamental frequency, HNR, jitter, and shimmer.

2. Speech analysis using the harmonic model

2.1. The harmonic model

The popular source-channel model of voiced speech considers glottal pulses resulting from the rhythmic opening and closing of vocal folds. These pulses are rich in harmonics, and are subsequently shaped by resonances of the oral cavity and the transfer function of the lip radiation. Given that the glottal pulse sequence is rich in harmonics, the resulting voiced sounds can be modeled with a harmonic model containing only non-zero Fourier components at harmonics of the period of the glottal pulses. The harmonic model is a special case of a sinusoidal model where all the sinusoidal components are assumed to be harmonically related; that is, the frequencies of the sinusoids are multiples of the fundamental frequency. This assumption arises from the harmonic nature of the speech signal and reduces the number of parameters in the general sinusoidal model (Stylianou, 1996). Stylianou (1996) introduced a harmonic plus noise model (HNM) for speech analysis and synthesis. The observed voiced signal in HNM is represented in terms of a harmonic component and a non-periodic component related to noise. Speech decomposition using a HNM is useful for applications in speech synthesis, voice conversion, speech enhancement, and speech coding.

2.2. Model expression

To review the model, we adopt the notations from Stylianou (1996). Let $\mathbf{y} = [y(t_1), y(t_2), \ldots, y(t_N)]^T$ denote the N speech samples in a voiced frame, measured at times $t_1, t_2, \ldots, t_N$. The samples can be represented with a harmonic model with an additive noise $\mathbf{n} = [n(t_1), n(t_2), \ldots, n(t_N)]^T$ modeled by a $\mathcal{N}(\mu, \sigma_n^2)$ as follows:

$$\begin{aligned} s(t) &= a_0 + \sum_{h=1}^{H} a_h \cos(2\pi f_0 h t) + b_h \sin(2\pi f_0 h t) \\ y(t) &= s(t) + n(t) \end{aligned} \tag{1}$$