Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 47–58

www.elsevier.com/locate/csl

Dempster-Shafer theory for enhanced statistical model-based voice activity detection☆

Tae-Jun Park, Joon-Hyuk Chang*

School of Electronic Engineering, Hanyang University, Seoul 04763, Republic of Korea

Received 16 November 2016; received in revised form 3 July 2017; accepted 3 July 2017
Available online 12 July 2017

Abstract

In this paper, we propose to combine the posterior probabilities of voice activity derived from different statistical model-based algorithms for enhanced voice activity detection. For this, the Dempster-Shafer (DS) theory of evidence is employed to represent and combine the different probabilities estimated by three different statistical model-based VAD algorithms including the Sohn's likelihood ratio test (LRT)-based method, smoothed LRT-based method, and multiple observation LRT-based method. By considering a generalization of the Bayesian framework and permitting the characterization of uncertainty and ignorance through the DS theory, the probability of an ignorant state is eliminated through the orthogonal sum of several speech presence probabilities, which results in the performance improvement when detecting voice activity. According to objective test results, it is discovered the proposed DS theory-based VAD method offers significant improvements over the conventional approaches.

Keywords: Dempster-Shafer theory; Voice activity detection; Likelihood ratio test

1. Introduction

Voice activity detection (VAD) is described as means of a finite state machine with at least two states, 'speech presence' and 'speech absence'. When speech is present at a low signal-to-noise ratio (SNR), the VAD becomes a key factor that significantly impacts the performance of the system for speech signal processing techniques. Over the years, various algorithms have been proposed for enhancing the performance of the VAD (ITU-T, 1996; Sohn et al., 1999; Cho et al., 2001; Ramírez et al., 2005; Chang et al., 2006; Shin et al., 2007; Kim et al., 2007; Kang et al., 2008; Kim and Chang, 2012; Fujimoto and Ishizuka, 2007; Ephraim and Malah, 1984; Hwang et al., 2016; Shin et al., 2010; 2008; Tan and Lindberg, 2010; Zhang and Wang, 2016; Zhang and Wu, 2013). Among them, the likelihood ratio test (LRT)-based decision rule for a set of hypotheses has been considered successful in many studies because of its high detection accuracy and implementation efficiency (Sohn et al., 1999; Cho et al., 2001; Ramírez et al., 2005). These methods initially assume the Gaussian statistical model for the LRT by adopting a decision-directed (DD) method for the given parameter estimation Ephraim and Malah (1984). The key to the successful design of

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

* Corresponding author.
E-mail address: jchang@hanyang.ac.kr (J.-H. Chang).

http://dx.doi.org/10.1016/j.csl.2017.07.001

the LRT-based VAD is to avoid relatively high numbers of detection errors in the offset region of the speech, so Sohn et al. (1999) addressed this issue to find a hangover scheme without an in-depth analysis. Then, the behavioral mechanisms of the LR were fully investigated in order to identify the unwanted phenomenon in the method of Cho et al. (2001) based on a smoothed LR to significantly improve the offset regions compared to the Sohn's method. Further, multiple observation (MO)-LRT was devised to incorporate the long-time information of speech activity into a decision rule based on a statistically optimum LRT that avoids the need to smooth the VAD decision but allows an inherent delay. However, the three methods mentioned above do not necessarily achieve the desired behavior in a variety of circumstances even though these schemes have been shown to dramatically increase the performance of the VAD at a specific condition.

In this paper, we present a novel VAD technique to apply Dempster-Shafer (DS) theory (Valente and Hermansky, 2007; Dempster, 1967; Shafer, 1976) to represent and combine the posterior probabilities derived from the LRT of three different statistical model-based VAD algorithms, including the Sohn's method Sohn et al. (1999), smoothed LRT Cho et al. (2001), and the MO-LRT scheme Ramírez et al. (2005). Since the DS theory enables the probability to have lower and upper bounds that the uncertainty in the evidence to be incorporated and a way to combine independent bodies of observation, an increase of the confidence in the given hypotheses of the VAD task can be clearly observed. For this, the posterior probabilities derived from the three different VAD algorithms are first converted into basic probability assignments (BPAs) in three different ways. For each conversion method, three BPAs are combined into an orthogonal sum, where the orthogonal sum is compared with a threshold value for the final VAD decision. For the performance evaluation, extensive objective experiments were carried out. It was found that the proposed DS theory-based VAD method outperforms conventional techniques under various noise conditions. The rest of this paper is organized as follows: Sections 2 and 3 describe the conventional statistical-based VAD algorithms and the proposed DS theory-based VAD algorithm. A series of simulations to evaluate the system performance are presented in Section 4. Then, Section 5 summarizes our contributions Fig. 1.

2. Conventional statistical model-based voice activity detection algorithms

In this section, we briefly explain statistical model-based VADs, including Sohn's LRT Sohn et al. (1999), smoothed LRT-based VAD Cho et al. (2001), and the MO-LRT-based one Ramírez et al. (2005). To this end adding the noise signal d to the undisturbed clean signal x yields a noisy signal y. Thus, the sum is denoted by the noisy speech signal y and taking a discrete Fourier transform (DFT) for the noisy speech y yields in the time-frequency domain:

$$Y(k,n) = X(k,n) + D(k,n) \tag{1}$$

where $Y(k, n)$, $X(k, n)$ and $D(k, n)$ denote the DFT coefficients of y, x, and d, respectively, with k and n denoting the frequency bin index ($k = 1, 2, \ldots, L$) and frame index ($n = 0, 1, \ldots$), respectively. To develop the detection algorithm, we formulate as

$$H_0 \quad : \quad Y(k,n) = D(k,n) \tag{2}$$

$$H_1 \quad : \quad Y(k,n) = X(k,n) + D(k,n). \tag{3}$$

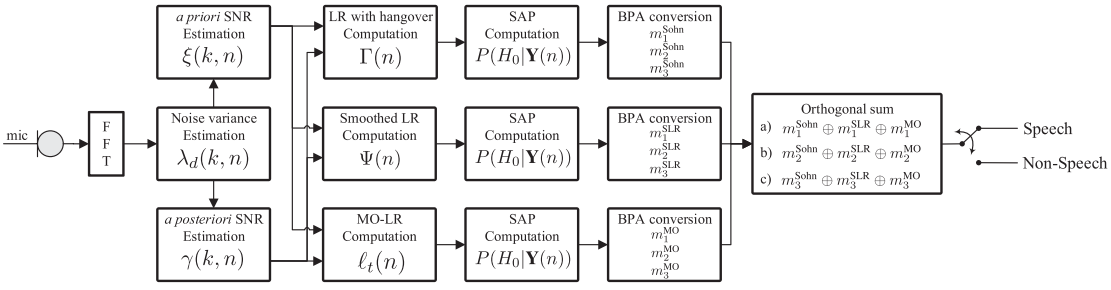

Fig. 1. Overall block diagram of the proposed DS theory-based VAD technique.