Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 403–416
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

MMSE estimation of log-filterbank energies for robust speech recognition

Anthony Stark, Kuldip Paliwal *

Signal Processing Laboratory, Griffith University, Nathan Campus, Brisbane QLD 4111, Australia

Received 15 June 2010; received in revised form 27 September 2010; accepted 3 November 2010
Available online 21 December 2010

Abstract

In this paper, we derive a minimum mean square error log-filterbank energy estimator for environment-robust automatic speech recognition. While several such estimators exist within the literature, most involve trade-offs between simplifications of the log-filterbank noise distortion model and analytical tractability. To avoid this limitation, we extend a well known spectral domain noise distortion model for use in the log-filterbank energy domain. To do this, several mathematical transformations are developed to transform spectral domain models into filterbank and log-filterbank energy models. As a result, a new estimator is developed that allows for robust estimation of both log-filterbank energies and subsequent Mel-frequency cepstral coefficients. The proposed estimator is evaluated over the Aurora2, and RM speech recognition tasks, with results showing a significant reduction in word recognition error over both baseline results and several competing estimators.

Keywords: Robust speech recognition; MMSE estimation; Speech enhancement methods

1. Introduction

State-of-the-art automatic speech recognition (ASR) can exhibit impressive recognition performance under laboratory conditions. Unfortunately, performance tends to degrade substantially when ASR is used in real world environments. This degradation is caused by acoustic model mismatch. Here, we use the term mismatch to describe any difference between the acoustic environment the ASR system was trained on, and the acoustic environment the ASR system is actually deployed in. Mismatch can include additive background noise, echoes, transmission channel effects, inter-speaker variability and intra-speaker variability. Taken together, such effects can rapidly reduce recognition accuracy to unacceptably low levels.

For this paper, we address the problem of additive background noise robustness. In the literature, several approaches have previously been proposed. Most fall under the following general categories: robust feature selection and extraction (Davis and Mermelstein, 1990; Hermansky, 1990), speech enhancement (Lathoud et al., 2005; Ephraim and Trees, 1991; Gemello et al., 2006; Hermus et al., 2007; Fujimoto and Ariki, 2000), model adaptation (Gales, 1995; Acero et al., 2000), model-based feature enhancement (Stouten, 2006; Moreno, 1996), missing feature theory (Raj and Stern, 2005; Cooke et al., 2000; Barker et al., 2000) and multistyle training (Deng et al., 2000).

In this paper, we focus on the problem of estimating robust Mel-frequency cepstral coefficients (MFCCs). In particular, we investigate the stochastic estimation of a clean speech MFCC vector from speech that has been corrupted with additive noise. Under the additive noise assumption, a noisy speech signal $y(n)$ is given by

$$y(n) = x(n) + d(n), \quad 0 \leqslant n < N, \tag{1}$$

* Corresponding author.
E-mail addresses: a.stark@griffith.edu.au (A. Stark), k.paliwal@griffith.edu.au (K. Paliwal).
URL: http://maxwell.me.gu.edu.au/spl/ (K. Paliwal).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.11.004

where $x(n)$ and $d(n)$ are the clean speech and noise signal, respectively. Since speech is often assumed to be quasi-stationary over short-time (20–40 ms) intervals, it is typically decomposed with framing. Here, the mth noisy speech frame can be given as

$$\boldsymbol{y}_m = \boldsymbol{x}_m + \boldsymbol{n}_m, \tag{2}$$

where $\boldsymbol{y}_m = [y(mS), y(mS+1), \ldots, y(mS+L-1)]^T$, L is the analysis frame length and S is the analysis frame shift. After discrete short-time Fourier transform (DSTFT) analysis (Rabiner and Schafer, 1978) of (2), we then have the following relationship

$$\boldsymbol{Y}_m = \boldsymbol{X}_m + \boldsymbol{D}_m, \tag{3}$$

where $\boldsymbol{Y}_m, \boldsymbol{X}_m, \boldsymbol{D}_m \in \mathbb{C}^{K\times 1}$ are the noisy speech, clean speech and noise spectral domain vectors (for the mth DSTFT analysis frame), respectively. For notational convenience, we drop the frame index m and dependence on this subscript is implicitly assumed henceforth.

Given the observed noisy speech vector, the goal of a minimum-mean-square error (MMSE) MFCC estimator is the determination of estimate $\hat{\boldsymbol{c}}$, where

$$\hat{\boldsymbol{c}} = E[\boldsymbol{c}|\boldsymbol{Y}], \tag{4}$$

where $\boldsymbol{c}$ is the clean speech MFCC vector, $E[\cdot]$ is the expectation operator and $\hat{\boldsymbol{c}}$ is the estimate that minimizes the mean-square-error to the true clean speech MFCC vector $\boldsymbol{c}$.

While the spectral domain noise distortion model (3) is straightforward, the estimation (4) is not. This is due to the highly non-linear relationship between spectral-domain speech and the MFCC vector. Given a spectral domain speech vector $\boldsymbol{Y}$, several intermediate variables must first be calculated: $\boldsymbol{e}$ – spectral energies, $\boldsymbol{E}$ – filterbank energies and $\boldsymbol{L}$ – log-filterbank energies. Fig. 1 shows the operations required for converting spectral-domain speech into an MFCC vector.

Instead of directly estimating the MFCC vector, we may focus our attention on one of the intermediate feature sets – namely log-filterbank energies. The MMSE log-filterbank estimate $\widehat{\boldsymbol{L}}$ is given by

$$\widehat{\boldsymbol{L}} = E[\boldsymbol{L}|\boldsymbol{Y}]. \tag{5}$$

where $\boldsymbol{L}$ is the clean speech log-filterbank energy vector. Since MFCCs and log-filterbank energies are linearly related, given $\widehat{\boldsymbol{L}}$ it is easy to find the MMSE MFCC estimate $\hat{\boldsymbol{c}}$

$$\hat{\boldsymbol{c}} = \boldsymbol{C}\widehat{\boldsymbol{L}}, \tag{6}$$

where $\boldsymbol{C}$ is the discrete cosine transform matrix. Thus, the core MFCC estimation problem now becomes a log-filterbank energy estimation problem. Unfortunately, a highly non-linear relationship persists between the spectral domain speech and its corresponding log-filterbank energy vector. Several strategies have been adopted in past literature to address this issue, including forced linearization of the noise model (Moreno, 1996 ; Stouten, 2006), numerical integration (of an analytically intractable model) (Erell and Weintraub, 1993) and development of simpler (and more tractable) noise distortion models (Yu et al., 2008; Indrebo et al., 2008). Each of the aforementioned methods is suboptimal in some manner, typically offering a trade-off between noise model simplification and computational tractability.

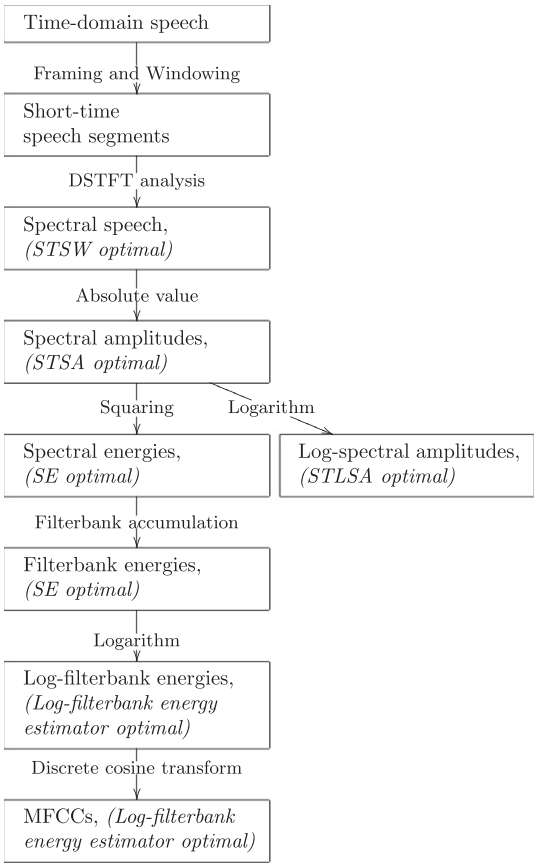

Fig. 1. Overview of the computation required for converting a time-domain frame of speech into an MFCC vector. MMSE optimality is achieved by several common spectral estimators at various intermediate stages of the MFCC derivation.

In many cases, a speech enhancement algorithm from the human listening domain is carried over to the machine recognition domain. Methods such as the short-time spectral amplitude (STSA) estimator and the short-time log-spectral amplitude estimator (STLSA) have commonly been used to reduce the effects of noise from MFCC features. However, the mathematical optimality of these estimators do not provide an exact match with the objectives of ASR – that is, reduction of error within the MFCC/log-filterbank domain. Fig. 1 highlights feature stages where the STSA, STLSA, short-time spectral Wiener (STSW) and short-time spectral energy (SE) estimators are optimal (in the MMSE sense). While none of the aforementioned estimators is strictly optimal (in the log-filterbank MMSE sense), they are all closely related. Because of this, we examine this class of estimators in greater detail, examining their relationship to an MMSE log-filterbank energy estimator.