Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 75 (2015) 62–75

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Reconstruction-based speech enhancement from robust acoustic features

Philip Harding, Ben Milner *

School of Computing Sciences, University of East Anglia, UK

Received 10 February 2015; received in revised form 13 August 2015; accepted 21 September 2015
Available online 9 October 2015

Abstract

This paper proposes a method of speech enhancement where a clean speech signal is reconstructed from a sinusoidal model of speech production and a set of acoustic speech features. The acoustic features are estimated from noisy speech and comprise, for each frame, a voicing classification (voiced, unvoiced or non-speech), fundamental frequency (for voiced frames) and spectral envelope. Rather than using different algorithms to estimate each parameter, a single statistical model is developed. This comprises a set of acoustic models and has similarity to the acoustic modelling used in speech recognition. This allows noise and speaker adaptation to be applied to acoustic feature estimation to improve robustness. Objective and subjective tests compare reconstruction-based enhancement with other methods of enhancement and show the proposed method to be highly effective at removing noise.

Keywords: Speech enhancement; Noise reduction; Adaptation; Sinusoidal model

1. Introduction

This work proposes a reconstruction-based approach to speech enhancement that aims to produce a noise-free signal. This moves away from conventional enhancement methods that use filtering to remove noise. Instead, the enhanced signal is reconstructed from a model of speech production and a set of acoustic features that are estimated from the noisy speech.

Historically, most approaches to speech enhancement are filtering methods implemented as analysis-modification-synthesis systems. Many filtering approaches have been proposed and have been categorised into spectral subtraction, Wiener filtering, statistical and subspace methods (Loizou, 2007). Spectral subtraction is the most simple and requires just a noise spectral estimate but is prone to musical noise and speech distortion and leaving residual noise (Boll, 1979; Hu et al., 2002). Wiener filtering produces higher quality speech, Loizou (2007, Ch. 11.2), although its implementation is more complex and requires an estimate of the SNR in each frequency bin. Iterative approaches, decision directed methods, nonnegative matrix factorisation and Gaussian mixture models (GMMs) have all been used to estimate the required *a priori* SNR (Scalart and Vieira-Filho, 1996; Mohammadiha et al., 2013; Hadir et al., 2011). The Wiener filter is the optimal, in the mean square error sense, linear estimator of the complex speech spectrum. Statistical methods extend this by identifying optimal non-linear minimum mean square error (MMSE) estimators of spectral magnitudes or log spectral magnitudes which give further improvement in speech quality (Ephraim and Malah, 1984, 1985). Extensions use probability density functions that better model the distribution of the speech spectra (Martin, 2005), and by taking into account speech presence uncertainty in estimating spectral amplitudes (Cohen, 2002). Subspace methods

* Corresponding author.

http://dx.doi.org/10.1016/j.specom.2015.09.011
0167-6393/© 2015 Elsevier B.V. All rights reserved.

transform the noisy speech into speech and noise subspaces (Hu and Loizou, 2003). Truncation of the vector space aims to retain elements containing speech and remove noise components, although retaining too few speech components oversmooths the speech, while retaining too many components leaves residual noise. Evaluation across a range of speech enhancement methods shows these filtering methods to be effective in improving speech quality but susceptible to the accuracy of noise and SNR estimates which can introduce unwanted artefacts into the enhanced speech such as musical noise, residual noise and distortion (Loizou, 2007, Ch. 11.2). Most filtering methods enhance the magnitude spectrum and combine this with the noisy phase, although some recent work has considered the importance of phase. Phase spectrum compensation is one example and adjusts the phase spectrum according to the noise magnitude spectrum and has been combined with MMSE spectral magnitude estimation (Stark et al., 2008; Paliwal et al., 2011).

An alternative to filtering the noisy speech is to reconstruct or synthesise a clean speech signal. This is motivated by a desire to reduce artefacts introduced by the filtering process. Reconstruction approaches can be loosely divided into those that reconstruct the speech using a model of speech production and those that use a corpus or inventory of clean speech segments to synthesise an enhanced speech signal.

The main challenge for model-based approaches is to obtain a set of noise-free speech parameters that can be applied to a model of speech production. The sinusoidal model has been used for speech enhancement where an initial set of model parameters is extracted from the noisy speech and then refined iteratively by Wiener filtering and smoothing (Jensen and Hansen, 2001). In Yan et al. (2008), noisy speech is first pre-cleaned and then decomposed into excitation and vocal tract components. A harmonic plus noise model (HNM) models clean excitation while formants are tracked using a combined Viterbi/Kalman filter, which are then combined using a linear predictive model. Chen et al. (2012) adopted a similar approach although uses a HNM to model the speech signal rather than the excitation. The noisy speech is again pre-cleaned and HNM analysis applied to extract fundamental frequency, spectral envelope and spectral gain with Kalman filtering tracking the parameters over time. The recent work in hidden Markov model (HMM)-based speech synthesis, Zen et al. (2009), has also been applied to speech enhancement (Carmona et al., 2013; Kato and Milner, 2014). These methods use a network of HMMs to decode noisy speech into a model and state sequence which is then input into an HMM-based synthesiser to output a clean speech signal.

Corpus-based approaches use a large database of clean speech and assume that noise varies more slowly than speech (Ming and Crookes, 2014). Given noisy speech, the speech corpus is searched for segments that when added to a stationary noise segment best resembles the noisy speech. Selected segments are then concatenated to form the enhanced signal. A related technique is inventory-style enhancement which utilises clean and noisy codebooks (Xiao and Nickel, 2010). The two codebooks are created from speech data taken from a single speaker and are matched as the noisy codebook is formed from the speech data but with noise added. During enhancement a hidden Markov model (HMM) finds the optimal sequence of codebook entries from the noisy codebook which in turn identifies waveform units that are concatenated to form the enhanced speech signal. A similar method uses two GMMs, one trained on noisy MFCCs and the other on clean MFCCs (Boucheron et al., 2012,). During enhancement the noisy input speech is matched to mixture components from the noisy GMM and then mapped to the clean GMM which outputs a stream of MFCC vectors that are inverted to form the enhanced signal.

The work presented here uses the reconstruction approach to speech enhancement which, given a sufficiently good speech model and a set of noise-free acoustic features, should produce speech that is free from residual noise and artefacts. This gives rise to two main challenges: (i) to find a sufficiently good model for speech reconstruction and (ii) to develop robust methods to estimate accurately noise-free acoustic speech features. Our approach is based upon a variant of the sinusoidal model and differs from previous approaches as no pre-filtering of the speech is needed before parameter estimation. Instead we propose an integrated statistical method that estimates the set of acoustic features needed for reconstruction within a single statistical framework. This relates to earlier work that reconstructed speech solely from a sequence of MFCC vectors within a distributed speech recognition (DSR) architecture using a sinusoidal model (Milner and Shao, 2007). Spectral envelope parameters were obtained by inverting the MFCC vectors while a maximum a posteriori (MAP) estimate of fundamental frequency was made from the MFCC vector. This generated good quality speech although was designed for clean speech input, was speaker-dependent and constrained to operate on 23-D MFCC vectors. Later work (Milner and Darch, 2011), also within a DSR architecture, considered statistical methods for estimating each acoustic speech feature separately from noisy MFCC vectors by including some noise compensation. The proposed work now uses a single statistical model to estimate the set of acoustic features and improves noise robustness by considering the effect of phase within a mismatch function which is applied using an unscented transform to adapt the clean model statistics to noisy speech. Furthermore, speaker adaptation is also applied to adapt model parameters to the speaker under test which removes the speaker dependence constraint of earlier systems. This set of robust acoustic speech features is then input into a variant of the sinusoidal model to reconstruct a speech signal and forms the proposed method of enhancement.

Section 2 examines speech production models for their suitability in reconstruction-based speech enhancement.