Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Speech Communication xxx (2015) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Speech enhancement using Bayesian estimation given a priori knowledge of clean speech phase

Q1 Sunnydayal Vanambathina*, T. Kishore Kumar[1]

Department of Electronics and Communication Engineering, National Institute of Technology Warangal, Warangal, Telangana 506 004, India

Received 5 May 2015; received in revised form 29 October 2015; accepted 9 November 2015
Available online xxx

Abstract

In this paper, STFT based speech enhancement algorithms based on estimation of short time spectral amplitudes are proposed. These algorithms use Maximum Likelihood (ML), Maximum a posterior (MAP) and Minimum mean square error (MMSE) estimators which respectively uses Laplace, Gaussian probability density functions (pdf) as noise spectral amplitude priors and Nakagami, Gamma distributions as speech spectral amplitude priors. The method uses a joint MMSE estimate of the clean speech amplitude and clean speech phase for a given uncertainty phase information for improved single channel speech enhancement. In the most of the speech enhancement algorithms, we only concentrate on the frequency domain amplitude of speech, but not on the phase of noisy speech since it may cause undesired artifacts. In this paper, a recent phase reconstruction algorithm is used to estimate the phase of clean speech. The reconstructed phase is treated as an uncertain prior knowledge when deriving a joint MMSE estimate of the Complex speech coefficients given Uncertain Phase (CUP) information. The proposed MMSE optimal CUP estimator reduces undesired artifacts and also gives satisfactory values between the phase of noisy signal and the estimate of prior phase. We evaluate all the above estimators using speech signals uttered by 10 male speakers and 10 female speakers are taken from TIMIT database. The proposed method outperforms other benchmark algorithms in terms of segmental signal to noise ratio (SSNR), short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ).

Keywords: ML estimator; MAP estimator; MMSE estimator; Laplace density; *von Mises* distribution; Nakagami distribution.

1. Introduction

In mobile communications, speech enhancement plays a very important role. The main goal of speech enhancement is to improve the quality and intelligibility which is degraded when the clean speech signal is corrupted by noise. In some of the traditional speech enhancement techniques (Krawczyk and Gerkmann, 2014), the input speech signal is divided into frequency bands which are processed separately and finally combined to get the output. For some long duration speech signals (e.g. vowels) frequency components are stationary while for some short duration speech signals (e.g. consonants) the frequency range is wide. It is difficult to find a trade-off between resolution in frequency and resolution in time, if the speech analysis is not adapted to the signal components.

Q2

In most of the noise reduction techniques (Hendriks et al., 2013), the modifications take place in the speech magnitude and there is no change in the noisy phase. Recently, some speech enhancement algorithms have shown that there may be improvements in speech enhancement if we know the phase of clean speech (Paliwal et al., 2011). The role of phase has been discussed in single channel enhancement techniques (Wang and Lim, 1982; Vary, 1985). The complex speech coefficients can be modeled as circular symmetric probability density function (PDF) (Erkelens et al., 2007). If we consider PDF as circular symmetric, the phase is uniformly distributed. In some of the speech enhancement algorithms (Wang and Lim, 1982), the noisy phase is replaced with clean speech phase. The phase of the clean speech can be reconstructed by using iterative STFT analysis and synthesis if and only if the clean speech magni-

* Corresponding author. Tel.: +91 9966380100.
E-mail addresses: sunny.conference@gmail.com
, sunnydayal45@gmail.com (S. Vanambathina), kishorefr@gmail.com (T.K. Kumar).
[1] Tel.: +91 8332 969 353.

http://dx.doi.org/10.1016/j.specom.2015.11.004

Please cite this article as: S. Vanambathina, T.K. Kumar, Speech enhancement using Bayesian estimation given a priori knowledge of clean speech phase, Speech Communication (2015), http://dx.doi.org/10.1016/j.specom.2015.11.004

tude is known (Griffin and Lim, 1984). The phase estimate is incorporated in Gerkmann and Krawczyk (2013) and Krawczyk et al., (2013) to improve Bayesian amplitude estimation.

The clean speech phase can be estimated by iteratively synthesizing and reanalyzing the clean speech magnitudes (Griffin and Lim, 1984). To implement these algorithms, we should know the clean speech magnitude *a priori*. If only the estimates are available, there may be chances of degradation in the enhanced speech. Recently there have been advances in iterative phase estimation (Sturmel and Daudet, 2011; Roux and Vincent, 2013; Mowlaee and Saeidi, 2013). The MMSE estimator of the clean speech spectral magnitude that uses both a parametric compression function in the estimation error criterion and a parametric prior distribution for the statistical model of the clean speech magnitude was proposed in Breithaupt et al. (2008). In Krawczyk and Gerkmann (2012), clean speech and the fundamental frequency of voiced speech is estimated. Using estimate of clean speech phase (Gerkmann and Krawczyk, 2013) instead of noisy speech phase for the reconstruction of clean speech introduces artifacts (Sturmel and Daudet, 2011; Krawczyk and Gerkmann, 2012). The proposed method addresses this problem by using ML, MAP and MMSE estimators.

Bayesian estimators like MMSE and MAP estimators are popular in estimating the clean speech coefficients.

For speech enhancement (Ephraim and Malah, 1984), short time spectral amplitude (STSA) of speech signal can be estimated and combined with short time phase of degraded speech for reconstructing the enhanced speech (Example Spectral Subtraction algorithm and wiener filtering). In "Spectral subtraction" algorithm, STSA is estimated as the square root of ML estimator of each signal spectral component whereas in wiener filtering, STSA estimator is obtained as the modulus of optimal MMSE estimator of each signal spectral component. Gaussian assumption is made in deriving these two STSA estimators. These two estimators are not optimal spectral estimators under the assumed statistical model. For deriving MMSE STSA estimator, the *apriori* probability distribution of speech and noise should be known.

The MMSE STSA estimator based on the statistical model was derived and compared with the wiener STSA estimator in Ephraim and Malah (1984). The estimator takes into account the uncertainty of speech presence in the noisy observations and estimates the complex exponential of the phase (Ephraim and Malah, 1984). In the reconstruction of enhanced signal, the complex exponential estimator is used in conjunction with MMSE STSA estimator. The MMSE complex exponential estimator does not affect STSA estimation and hence the noisy phase can be used for reconstruction. At high signal to noise ratios (SNR), MMSE estimator and wiener amplitude estimator converges. MMSE estimator is derived under the assumption that *a priori* SNR and noise variance are known. Wiener and MMSE estimators are more sensitive to the underestimate of *a priori* SNR than its overestimate. In wiener estimator, residual mean square error decreases as the *a priori* SNR overestimates.

The ML estimation is used to estimate an unknown parameter of a given PDF, when *a priori* information is not available. MMSE estimator or Wiener estimator gives similar enhanced quality speech when *a priori* is estimated by ML estimator. "Musical noise" increases as input SNR decreases. Enhanced speech quality obtained by MMSE estimator with either ML *a priori* SNR or "decision-directed" *a priori* SNR are similar. Wiener estimator with "decision-directed" approach yields more distorted speech than MMSE estimator with "decision-directed" approach. At high SNRs, wiener estimator and MMSE estimators are similar but at low SNRs, MMSE estimator gives less mean square error (MSE). Measured phase will not provide any useful information in the suppression of noise. In the comparison of suppression rules of Wiener filtering and ML algorithms, the gain functions are similar at high SNRs. As SNR decreases there is more increase in gain in ML than the Wiener estimator (Robert Mcaulay and Malpass, 1980). Since at low SNRs, "most likely" corresponds to noise alone, the effect of residual noise should be reduced. At large SNRs, "most-likely" means speech present and so the speech envelope can be extracted using ML estimator.

The assumption of Gaussian prior is made for clean speech Discrete Fourier Transform (DFT) coefficients in Ephraim and Malah (1984), Ephraim and Malah (1985), and Cohen (2001). The assumption holds asymptotically for long duration analysis frames. In this case, the span of signal correlation is shorter than DFT size. The assumption may hold for DFT coefficients (real and imaginary parts) of noise but does not hold for DFT coefficients of speech (real and imaginary parts), since the speech coefficients are estimated using short duration windows (20–30 ms) (Chen and Loizou, 2007). To resolve this shortcoming, non-Gaussian distributions (Laplacian and Gamma PDF) have been employed (Chen and Loizou, 2005; Hendriks and Heusdens, 2010).

The assumption of DFT coefficients as Gamma PDF provides better fit to the experimental data and also provides smaller Kullback divergence when compared with Gaussian distribution (Lotter and Vary, 2003).

Rician distribution is approximated by the Nakagami distribution (Xie and Zhang, 2014) to estimate speech spectral magnitude. The approximation is widely used in wireless communication (Wang and Lea, 1998) since Rician distribution contains a modified Bessel function which is difficult to solve and also minimizing the cost function using this distribution is difficult. The Nakagami distribution prior preserves speech spectral components at the expense of a larger number of spurious spectral peaks. The Gamma prior suppresses weaker spectral components. In the noise dominated regions of the spectrogram, Nakagami distribution prior results in smoother spectral peaks and hence, the residual noise of the enhanced sentence is more uniform.

An MMSE estimator was developed with speech DFT coefficients modeled by Gamma distribution (Martin, 2005). In Lotter and Vary (2004) MAP estimator was shown to outperform Ephraim–Malah estimator with Laplace DFT coefficients. MAP magnitude estimation considering speech coefficients as Gamma and Rice distribution was proposed in Dat et al. (2004). MAP based speech enhancement (Dat et al., 2005), modeling the speech spectral coefficients with Generalized Gamma, fitted the