ISSN No. 0976-5697

Volume 5, No. 4, April 2014 (Special Issue)

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Performance Evaluation of Various Speech Enhancement Techniques

Prathamesh V. Phadke[1], V.M. Thakare[2], R.N. Khobragade
S.G.B.A.U, Amravati Maharashtra India
phadkepv@gmail.com[1], vilthakare@yahoo.co.in[2], rnkhobragade@gmail.com[3]

Abstract: Temporal dynamics and speaker characteristics are two important features of speech that distinguish speech from noise. In this paper, aim is to propose a method to maximally extract these two features of speech for speech enhancement. This can reduce the requirement for prior information about the noise, which can be difficult to estimate for fast-varying noise. Given noisy speech, the new approach estimates clean speech by recognizing long segments of the clean speech as whole units. In the speech recognition, clean speech sentences, taken from a speech corpus, are used as examples. Matching segments are identified between the noisy sentence and the corpus sentences. The a priori signal-to-noise ratio (SNR) plays an important role in many speech enhancement algorithms. It may be used with a wide range of speech enhancement techniques, such as, e.g., the minimum mean square error (MMSE) (log) spectral amplitude estimator, the super Gaussian joint maximum a posteriori (JMAP) estimator, or theWiener filter. Also, Discrete cosine transform (DCT) has been proven to be a good approximation to the Karhunen–Loeve Transform (KLT) and has similar properties to the discrete Fourier transform (DFT). This Paper suggests a better energy compaction capability which is advantageous for speech enhancement.

I. INTRODUCTION

A speech signal has two distinct features**: its temporal dynamics**, subject to acoustic, lexical, and language constraints, and its **speaker characteristics**. These two features distinguish a speech sentence from non-speech noise, and from other speakers' sentences. a method to maximally extract these two features of speech for retrieving speech from noise, including crosstalk interference[1]. Apart from that, it is also important to maintain the naturalness of the enhanced signal since artifacts such as musical tones will be perceived as particularly disturbing. The computation of spectral weighting rules in speech enhancement are often driven by the a posteriori and the a priori signal-to-noise ratio (SNR). However, the performance of most weighting rules is dominantly determined by the a priori SNR, while the a posteriori SNR acts merely as a correction parameter in case of low a priori SNR[2]. **Speech enhancement** can be performed both in the time domain and the frequency domain. Time domain filters include those utilizing **finite impulse response (FIR)** and **infinite impulse response (IIR) filters**, linear predictive coefficients (LPCs) , **Hidden Markov model** (HMM) etc. Transform domain filters are those which calculate the transform coefficients first followed by the enhancement process. For single-channel speech enhancement, a number of transform- based algorithms have been investigated in the past. Among these, DFT-based algorithms are the most active[3].

A. Background:

In optimal estimation, parametric statistical models such as Gaussian, Gamma, Laplacian, or super Gaussian have found use in representing the probability distribution of the speech discrete Fourier transform (DFT) coefficients or spectral amplitudes. **Data-driven models**, such as vector quantization (VQ) codebooks or **Gaussian mixture models (GMMs**), have also been used to provide the speech priors required in optimal\ estimators [1]. The most widespread approach to determine the a priori SNR estimates is the decision-directed (DD) estimator. The a priori signal-to-noise ratio (SNR) plays an important role in many speech enhancement algorithms. In a data-driven approach to a priori SNR estimation it may be used with a wide range of speech enhancement techniques, such as, e.g., the minimum mean square error (MMSE) (log) spectral amplitude estimator, the super Gaussian joint maximum a posteriori (JMAP) estimator, or theWiener filter [2].

Another important area of speech enhancement in Fourier domain is the **statistical estimation problem** based on two statistical models of speech signal and noise signal. In **the pitch synchronous overlap add** (PSOLA) method is applied in the time domain and it renders the proposed algorithm to be able to control the value of the synthesized pitch and the duration of the synthesized signal. The PSOLA method can also be applied in other domains such as frequency domain. The Fourier transform is applied on the pitch synchronous segments and the resulting spectra are approximated by a pattern of zeros and poles in order to obtain the pitch synchronous representation for analyzing the voiced sounds[3].

B. Previous work:

Previous research has revealed the importance of imposing cross-time spectral constraints in improving speech enhancement quality. In past single-channel techniques for speech enhancement typically include optimal filtering and optimal estimation. In optimal filtering, no specific knowledge about the speech is assumed, except its independence of the noise; examples include spectral subtraction or Wiener filtering. In optimal estimation, a priori knowledge of the probability distribution of the speech is assumed and this is used to derive the estimators, for example, minimum mean-square error (MMSE), maximum a posteriori (MAP) or perceptually weighted Bayesian spectral estimators. In optimal estimation, parametric statistical models such as Gaussian, Gamma, Laplacian, or super Gaussian have found use in representing the probability distribution of the speech discrete Fourier transform (DFT) coefficients or spectral amplitudes. Data-driven models, such as vector quantization (VQ) codebooks or Gaussian mixture models (GMMs), have also been used to provide the speech priors required in optimal estimators [1].

In earlier an alternative, proposed **a non-causal a priori SNR estimator**. With a look ahead of a few frames it is capable of discriminating between speech onsets and irregularities in the a posteriori SNR corresponding to noise only, resulting in less transient distortion and less musical tones. Nevertheless, because of the non-causal estimation, it can be used only in applications that can tolerate additional delay. Besides spectral amplitude weighting rules, the data-driven approaches also show their strengths in the field of noise power spectral density (psd) estimation. Noise psd estimators have been evolved from voice activity detection (VAD)-driven approaches to minimum statistics and the **improved minima controlled recursive averaging** (IMCRA) method. Recently, Erkelens and Heusdens proposed a data-driven approach to iterative training of a noise psd weighting rule optimized under **the minimum mean square error (MMSE)** cost function. The noise psd estimate is then achieved by multiplying the noisy power spectrum with the resulting noise psd weighting rule[2].

Apart from that previous **DFT-based algorithms** only attempt to filter the spectral magnitude while leaving the noise corrupted phase information intact, since it has been reported that the best estimate of the phase is the corrupted one itself. Since such no action actually results in an upper bound on the maximum possible improvement in signal-to-noise ratio (SNR), DCT can therefore achieve a higher upper bound than DFT. The last advantage is very straightforward based on the transform properties. DFT only produces about half the independent spectral components as the other half are complex conjugates, while DCT produces fully independent spectral components. Based on these advantages, many practical work also prove that DCT is an acceptable alternative to the discrete Fourier transform (DFT) for speech enhancement. Pitch synchronous analysis is also an attractive topic and can be traced back to 1950s. It has been earlier used in numerous speech signal processing systems such as speech analysis/synthesis system, prosody modification system and speech recognition system . The basic idea of pitch synchronous processing is to firstly divide the speech signal into pitch periods for the voiced sounds and into pseudo pitch periods for unvoiced sounds[3].

C. Existing methodology:

Corpus-Based Approach to speech enhancement from Non-stationary noise is used to focus an approach aiming to maximally extract the two important features of speech—temporal dynamics and speaker-class characteristics for its separation from noise .Also, The subspace approaches have been used in speech enhancement, which project a noisy signal onto two subspaces: signal and noise; noise reduction is achieved by retraining only the signal-subspace projection, usually modified by filtering or speech prior[1].

Data-driven approach to a priori SNR Estimation approach focuses on the a priori SNR as a nonlinear function of the Data Driven contributing components, instead of using the weighted linear combination to reduce transient distortions as well as musical tones. For this purpose data-driven speech enhancement using either neural networks or simple lookup tables to compute an a priori SNR estimate[2].

DCT-Based Speech Enhancement System With Pitch Synchronous Analysis focuses its attention speech enhancement because of its excellent energy compaction property which is comparable with KLT but with a more efficient computational load. It is also a Fourier-related transform which only uses real numbers instead of the complex ones used in DFT. It also utilizes this pitch synchronous representation and applies Wavelet transform on it to obtain a new representation of pseudo-periodic signal in terms of a regularized oscillatory component and fluctuations[3].

II. ANALYSIS AND DISCUSSION

Current single-channel techniques for speech enhancement typically include optimal filtering and optimal estimation. In optimal filtering, no specific knowledge about the speech is assumed, except its independence of the noise; examples include spectral subtraction or Wiener filtering. In optimal estimation, a priori knowledge of the probability distribution of the speech is assumed and this is used to derive the estimators, for example, **minimum mean-square error (MMSE)**, **maximum *a* posteriori (MAP)** or perceptually weighted Bayesian spectral estimators . In optimal estimation, parametric statistical models such as Gaussian, Gamma, Laplacian, or super Gaussian have found use in representing the probability distribution of the speech discrete Fourier transform (DFT) coefficients or spectral amplitudes. Data-driven models, such as vector quantization (VQ) codebooks or **Gaussian mixture models (GMMs),** have also been used to provide the speech priors required in optimal\ estimators[1].

Recently, several data-driven techniques have been published showing the potential of this paradigm in the context of **speech enhancement**. The researchers have published various data-driven weighting rules, optimized for a specific acoustic environment of interest. By employing clean speech and noise training data, frequency-individual spectral amplitude weighting rules are trained as a function of a given **a posteriori SNR**, as well as **a priori SNR**, which is computed through a slightly modified decision-directed approach. Following a similar data-driven paradigm, Erkelens et al also proposed **a data-driven weighting rule**, trained under white noise with known spectral variance using a bias-compensated decision-directed a priori SNR estimation. Besides spectral amplitude weighting rules, the data-driven approaches also show their strengths in the field of noise **power spectral density (psd)** estimation[2].

An outline of the new data-driven approaches to a priori SNR estimation follows subsequently in **Data-Driven A Priori SNR Estimation**. In Data-Driven A Priori SNR Estimation, the performance is evaluated in comparison to the conventional decision-directed approach. In that the goal of speech enhancement is to compute an estimate of the clean speech signal. Following this work, it also utilizes the pitch synchronous representation and applies Wavelet transform on it to obtain a new representation of pseudo-periodic signal in terms of a regularized **oscillatory component and fluctuations**. This representation offers several scales for analyzing the fluctuations which is superior to Fourier representation with only one scale [3].

III. PROPOSED METHODOLOGY

A. *Modeling Long-Range Temporal Dynamics Of Speech and Identifying Matching Segments With Large Continuities:*

It consisting of prerecorded clean speech sentences by various speakers, to provide the required free-speaker and free-text acoustic, lexical, and language constraints for the target speech. A reasonably sized speech database, as normally used to develop HMM systems for large-vocabulary speaker-independent speech recognition, could suit the purpose. Each sentence in the corpus will serve, simultaneously, as an acoustic model of a speech process that may be partly or completely realized in the target sentence, and as a text-dependent model of the acoustic characteristics of the target speaker class.

Now, let $y = \{y_t : t = 1, 2, . . . , T\}$ be a noisy test sentence with frames and being the frame at time . In this system, the problem of speech enhancement can be stated as identifying for each noisy frame a matching corpus frame , such that the underlying target speech frame can be reconstructed using the clean corpus frame modeled by the Gaussian $g(x|m_{x,i})$.

B. *Noise Compensation and Neural Network A Priori SNR Estimator:*

Consider noise compensation without assuming specific knowledge about the noise.It achieve this by combining multi-condition model training and optimal feature selection. So call the method missing-feature based noise compensation, which has been studied previously within the HMM and GMM frameworks for robust speech and speaker recognition.

For the proposed data-driven a priori SNR computation, both neural networks for speech presence and absence are required to be trained beforehand. Having the clean speech spectra available in training, the so-called ideal a priori SNR can be computed as with the estimated noise psd taken instead of the real one. This reduces the mismatch between training and test, and therefore improves the performance of the data-driven a priori SNR estimator.

IV. RESULTS

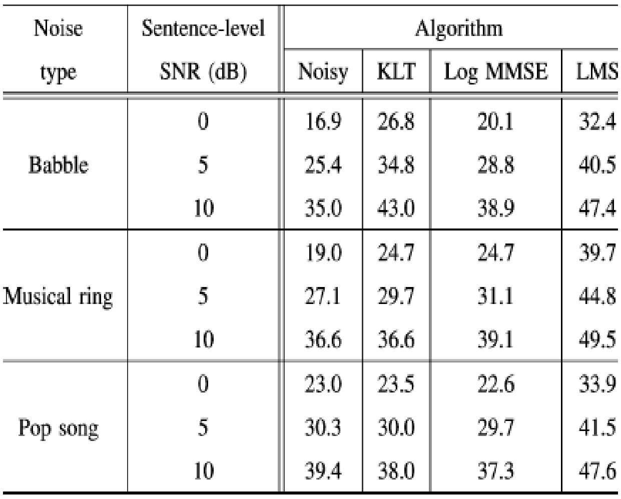
A. *LMS Algorithm*:

LMS algorithm consists of three core components: multicondition noise compensation with simulated noise, optimal feature selection to reduce the compensation mismatch, and estimation based on the longest matching segments to increase noise immunity. These three components combined offer robustness to non-stationary noise without assuming noise information. Using one test noise, musical ring, as an example, it studied the impact of each of the three components in terms of improving the objective measures. The second reduced algorithm has no feature selection (noted as "-FS"), which therefore compares the *full* set of time–frequency features between the noisy test sentence and the corpus sentences with simulated noise, to identify the longest matching segments. The last reduced algorithm (noted as "1-frame seg") has both noise compensation and feature selection as in the LMS algorithm, but does not search the longest matching segments for the estimation.

B. *Results of Speech Enhancement From Noise:*

The main issue is that it compares the new LMS algorithm with the conventional KLT, Log MMSE, MBand, and Wiener filtering algorithms for enhancing the TIMIT sentences from the three types of noise: babble, musical ring, and pop song. The comparisons include three objective quality measures—segmental SNR, logspectral distance, and phone identification accuracy—and informal subjective listening tests.

Table: 1



Noise type	Sentence-level SNR (dB)	Algorithm			
		Noisy	KLT	Log MMSE	LMS
Babble	0	16.9	26.8	20.1	32.4
	5	25.4	34.8	28.8	40.5
	10	35.0	43.0	38.9	47.4
Musical ring	0	19.0	24.7	24.7	39.7
	5	27.1	29.7	31.1	44.8
	10	36.6	36.6	39.1	49.5
Pop song	0	23.0	23.5	22.6	33.9
	5	30.3	30.0	29.7	41.5
	10	39.4	38.0	37.3	47.6

C. *Windowing Function:*

In signal processing, if a signal is to be observed over a finite duration, then a window function has to be applied to truncate this signal. The simplest window function is the rectangular window which causes the well-known problem, spectral leakage effect. That is, if there are two sinusoids with similar frequencies, leakage interferes with one buried by the other. If their frequencies are dissimilar, leakage interferes when one sinusoid is much weaker in amplitude than the other. The main reason is that the rectangular window represented in the frequency domain has strong side-lobes where the first side-lobe is only around 13 dB lower than the main lobe.

D. *Evaluation Metrics:*

The proposed ATSA technique is evaluated using two objective measures, segmental SNR (SegSNR) measure and perceptual evaluation of speech quality (PESQ) measure. Since SegSNR is better correlated with mean opinion score (MOS) than SNR as indicated by and is easy to implement, it has been widely used to qualify the enhanced speech.

a. *Pros:*

a) The subspace approaches have been used in speech enhancement, which project a noisy signal onto two subspaces: signal and noise; noise reduction is achieved by retraining only the signal-subspace projection, usually modified by filtering or speech prior.

b) All the techniques require prior knowledge about the noise, typically, the noise variance or power spectral density, or the instantaneous signal-to-noise ratio (SNR).

c) It have also been employed in many other areas for de-noising noisy measurements.

d) It provides a significantly higher energy compaction capability

e) It is a real transform without phase information

b. *Cons:*

a) Because of the non-stationary nature of the speech signal, most current enhancement algorithms operate on a frame-by-frame basis. Many algorithms ignore the temporal constraints between adjacent speech frames.

b) Without context, and without specific knowledge about the noise, it can be difficult to separate the speech from noise in the duration of a frame.

c) Apart from segmental SNR improvement, it is also reported that data-driven noise psd estimation not achieves faster noise tracking than the non-data-driven approaches, even in a non-stationary noise environment.

d) The variable shift used is only applicable to voiced speech signals, and not for unvoiced/silence parts of the speech signal, the algorithms falls back to the standard time-shift for the analysis window.

e) The main drawback of KLT-based algorithms is the high computational complexity.

c. *Applications* :

Reducing noise in song, music, and used in terms of crosstalk speech communication. Audio Applications, mobile applications and used in hearing devices.

d. *Future Scope:*

In the future the possibility of combining the conventional noise estimation algorithms, which are effective in tracking slow-varying noise, into the LMS algorithm, to improve the algorithm for predictable noise while retaining robustness to unpredictable noise

The proposed techniques can be integrated into a complete system named adaptive time-shift analysis **(ATSA)** speech enhancement system which produces good quality enhanced speech. Two objective measures, segmental SNR and PESQ are utilized to evaluate the proposed ATSA system. All the results show that significant improvements can be obtained by the proposed techniques.

V. CONCLUSION

In this a new approach to speech enhancement assuming a lack of prior information about the noise. This assumption applies to heavily non-stationary noise that can be difficult to predict with conventional noise estimation algorithms. The new approach, called LMS, aims to maximally extract two important features of speech-temporal dynamics and speaker characteristics-for retrieving speech from noise. It achieves this through recognition of long segments of the target speech as whole units.

Also, in this, two data-driven approaches to a priori SNR estimation. They each consist of two estimators, one for speech presence and one for speech absence, being implemented either as neural networks. Both estimators are trained with white noise, employing the two contributing SNR components of the conventional decision-directed approach. The neural networks approach outperforms the decision-directed a priori SNR estimator both in terms of speech distortion and noise attenuation.

In traditional **DCT-based noise reduction algorithms**, the observed speech signal is divided into fixed overlapping frames and transformed into DCT domain.

VI. REFERENCES

[1]. Ji Ming, Member Ramji Srinivasan, Danny Crookes, “A Corpus-Based Approach to Speech Enhancement From Nonstationary Noise”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 4,P.P. 822-837, MAY 2011.

[2]. Suhadi Suhadi, Carsten Last, and Tim Fingscheidt, “A Data-Driven Approach to A Priori SNR Estimation”, : IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 1, P.P.186-196 JANUARY 2011.

[3]. Huijun Ding, Ing Yann Soon, and Chai Kiat Yeo, “A DCT-Based Speech Enhancement System With Pitch Synchronous Analysis”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING VOL. 19, NO. 8, P.P.2614-2624 NOVEMBER 2011.