Modified Approach for Speech Enhancement Techniques

Mr. Prathamesh V. Phadke Dr. V. M. Thakare Dr.Mrs. Sujata N. Kale Mr. R.N.Khobragade

Abstract :- **Speech is the fundamental, most effective, reliable and common medium to communicate in real time systems. There are so many applications of speech still to be far from reality just because of lack of efficient and reliable noise removal mechanism and techniques for preserving or improving the intelligibility for the speech signals. In this paper attempt has been stepped towards surveying the methodologies for soft computing based speech recognition techniques for speech enhancement in multimedia applications. Because of their many applications and their relative ease of implementation, single-channel speech enhancement algorithms have received much attention. As a consequence, a vast amount of publications on estimation procedures and their implementation in noise reduction systems exists. However, there has been little systematic research on the theoretic performance of such estimators. Missing data techniques (MDTs) have been widely employed and shown to improve speech recognition results under noisy conditions. This paper presents a new technique which improves upon previously proposed sparse imputation techniques relying on the least absolute shrinkage and selection operator (LASSO). LASSO is widely employed in compressive sensing problems. However, the problem with LASSO is that it does not satisfy oracle properties in the event of a highly collinear dictionary, which happens with features extracted from most speech corpora.**

Keywords : **speech recognition techniques, Speech Enhancement**

I. INTRODUCTION

Speech is the fundamental, most effective, reliable and common medium to communicate in real time systems. In market due to advancement in technology many speech communication applications based devices is available, they are cheaper and easily available. However, undesired noises in environment cause undesired effects in real time speech processing systems. Human communications and intelligent machines are suffers from the degraded performance in which they takes decision based on what it receives as a speech. The speech enhancement is useful for storage and transmission of speech data, also it improves speech recognition based system performance where accurate identification of words and sentences can provide automation in most of the human-machine or machine based interface[1].In Single-channel speech enhancement, it is widely used in applications such as mobile communications or hearing devices. A vast amount of literature has been published on spectral-domain noise reduction algorithms for noisy speech signals, the best known being the Wiener filter and both the short-time spectral amplitude (STSA) and the log-spectral amplitude (LSA) estimator by Ephraim and Malah[2]. Missing data/feature techniques (MDTs) have been proposed for noisy signal conditions to compensate for unreliable components of features corrupted by noise. Missing data techniques have been employed in statistics long before its adoption into the speech processing field for automatic speech recognition (ASR). In addition to speech processing, techniques for data imputation have also been employed in many other areas for de-noising noisy measurements[3]. Applications of audio and speech processing include many well-reviewed algorithms for estimating the fundamental frequency of monophonic speech and music signals. In the case of polyphonic signals, it is more difficult to successfully estimate each of the fundamental frequencies, as reflected by the dearth of existing methods addressing this problem[4]. Due to the imperfection of speech acquisition and transmission system, speech is often corrupted by noise. The contamination on the speech not only affects its audio quality but also precludes many further high-level speech processing tasks such as speech signal coding and recognition[5]. Also, spoken language communication is a fundamental factor in quality of life, but as many as 1.3% of the population cannot use natural speech reliably to communicate, especially with strangers[6]. Now a days embedded speech recognition systems are becoming more important with the rapid development of handheld portable devices. However, only a few products are yet available due to the high chip costs[7]. In speech enhancement teleconferencing capability is an integral part of modern communication networks. It facilitates group collaborations efficiently and at low costs. A key technical challenge for a teleconferencing system is the ability to acquire high-fidelity speech while keeping speakers[8]. In addition to this Automatic speech recognition (ASR) is an enabling technology for a number of important information processing applications in the realm of human language technology[9]. ASR has made noticeable progress since the days it was considered a domain of ‘‘mad inventors or untrustworthy engineers’’ and it quite likely pushed signal information processing further than the processing of any other sensory signal[10].

II. BACKGROUND

There are number of speech recognition systems exists, some of them integrated into task specific applications. In practical applications a robust Mandarin Speech Recognition system using neural networks applied to multimedia interfaces performs better. The speech recognition used in a multimedia speech therapy system for different problems and different ages. In addition to recording the voice and analyzing the recorded spoken signal, speech recognition performs identification of speech irregularities and tracking the patient progress using time frequency analysis and neural network techniques. Feature extraction and coding stage reduces the dimensionality of the input vector and maintain discriminating power of the signal. The feature extraction because the number of training and test vector needed for the classification problem grows with the dimension of the given input. Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) are the most widely used methods for feature extraction[1]. The performance of the speech enhancement system will critically depend on the speech estimator itself, the SNR estimator, and on their joint temporal dynamics. The combination of the STSA estimator and the decision-directed approach is frequently used as a reference system in speech enhancement evaluations. However, to demonstrate, there is still a need for a more thorough understanding of the nonlinear, dynamic effects, and also a generalization to other speech estimators. In fact, the results cannot be generalized to other estimators such as the Wiener filter or the LSA estimator[2]. Sparse representation techniques have also been used in the realm of MDT, attempting data reconstruction under the assumption that the signal can be reconstructed by a sparse representation from a dictionary. Sparse representation techniques and compressive sensing techniques (where the dictionary obeys the restricted isometry hypothesis) have been used widely in applications including phonetic classification in speech processing and also image processing and medical imaging[3].

III. PREVIOUS WORK DONE

Speech enhancement can boost up the performance of speech recognition systems by keeping low word error rate. There are number of speech recognition systems exists, some of them integrated into task specific applications. In practical applications a robust Mandarin Speech Recognition system using neural networks applied to multimedia interfaces performs better. The speech recognition used in a multimedia speech therapy system for different problems and different ages. In addition to recording the voice and analyzing the recorded spoken signal, speech recognition performs identification of speech irregularities and tracking the patient progress using time frequency analysis and neural network techniques [1]. Analyze estimators for both the complex speech coefficients and their amplitude. For low SNR conditions, the SNR estimate in combination with a spectral shows a strong dependency on the type of estimator and its parameters. In Smoothing Properties of the Decision-Directed Approach In Low SNR Conditions as describe by decision-directed SNR-estimator recursive averaging of an observation of the constant is generally described as the smoothing constant like where it is evaluated in connection with the STSA estimator[2]. There have been a large number of works pertaining to the topic of MDT and imputation in the speech processing field. For example, two different statistical methods to infer the unreliable speech data. The first is marginalization, where the likelihood of the incomplete data vector is computed. Using x_r and x_u to denote the reliable parts and the unreliable parts of the feature vector respectively, the method allows computation $p(x_r \; C$ of instead of $p(x_r, x_u \; C)$, where C represents the states in a hidden Markov model (HMM). The second method is to compute the distribution of the unreliable segments of the feature vector instead of the likelihood of the data present. Experimental evaluation on the TIDigits corpus with non-stationary (car/helicopter/factory) noise corruption showed that with these proposed techniques, the performance is much better than the original performance before imputation[3].

IV. EXISTING METHODOLOGIES

The feature extraction because the number of training and test vector needed for the classification problem grows with the dimension of the given input. Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) are the most widely used methods for feature extraction. MFCC preferred over LPC because it is less prone to noise. The spectral signal output of speech analysis converted to activity signals on the auditory nerve using neural transduction method. Then activity signal converted into a language code within the brain, and finally message understanding is achieved.

i) Mean a Priori SNR Estimate in Low SNR conditions:

For the decision-directed recursion, an approximation of the mean estimate $\xi_k = \mathrm{E}\{\xi_k(l) \quad \mathcal{H}_k^0\}$ during low SNR conditions can be derived. During speech absence and in the case of Gaussian noise, an approximation of the mean a priori SNR in connection with the MOSIE estimator is [2]

$$\bar{\xi}_k = \frac{(1-\alpha)\mathrm{e}^{-1}}{1-\alpha R(\mu, \beta)}$$

ii) Mean Noise Suppression in Low SNR Conditions:

The mean value ξ_k from can be used to obtain a measure for the average degree of noise suppression that is possible using a noise suppression framework based on in combination with the decision-directed SNR estimation[2].

iii) Multiplicative Soft-Gain and its Influence on the Smoothing Effect

In the investigation of the smoothing effect is done only for the version of the STSA estimator that does not consider the conditional speech-presence probability as an additional, so-called soft-gain factor in the computation of the output amplitude [2].

V. ANALYSIS AND DISCUSSION

There are various types of advanced speech enhancement algorithms and they can be classified in main three categories, namely; filtering/estimation based noise reduction, beam forming and active noise cancellation (ANC) techniques[1].

i) "Colorless" Residual Noise of the STSA Estimator Using Soft-Gain:

The STSA estimator that considers speech-presence uncertainty is said to produce a residual noise that is also "colorless" and does not contain musical noise. Therefore, the smoothing is reduced, this obviously does not lead to an increase in the amount of musical noise. In order to clarify this contradicting observation, an analysis of the “colorless” residual noise that can result from the STSA estimator. The analysis is also meant to clarify the term “colorless” as it is often mentioned in connection with the noise reduction framework. Now show that a limitation of the a priori SNR estimate towards values $10\log_{10}\left(\xi_k(l)\right) \geq \xi_{\min,\mathrm{db}}$ in a preset range has a strong influence on both the amount of musical noise and the naturalness of the output signal[2].

It is analyzed both the estimate $\widehat{\xi}_k(l)$ and $\widehat{A}_k(l)$ the final filter output from either for conditions where the SNR continuously stays low and where typical noise amplitudes of a stationary Gaussian noise are observed. In this, it will analyze the noise reduction framework for two further cases: that of speech onsets and that of statistical outliers in the random noise signal. Note that in the case of, e.g., babble noise such outliers are far more likely than in the case of Gaussian noise. Further, the way these outliers are processed can have a great influence on the amount of musical noise in the processed signal. The two above-mentioned cases are very similar, as an outlier in the random noise signal cannot be distinguished from an onset of speech when only a single spectral bin K is observed.

ii) **Signal Reliability Masks:**

Most works in speech processing MDT define some sort of signal reliability mask for the mel-frequency log-energy coefficients (the popularly used signal representation), which is a matrix the size of the original feature vectors with entries containing 1 to mean that the feature component is reliable and 0 to mean domination by noise.[3]

$$M(k,t) = \begin{cases} 1, & \text{if } \frac{S(k,t)}{N(k,t)} > \lambda_{\text{SNR}} \\ 0, & \text{otherwise} \end{cases}$$

VI. Proposed Methodology

1)Speech Recognition System and construction of Representation of Test Utterances:

Speech recognition system is used as intelligence home in personal communication applications. It is also used in banking systems. It consists of four main building blocks speech analysis, feature extraction, language translation and message understanding. Speech analysis stage consists of noise removal, silence removal and end point detection. End point detection and removal of noise, silence is required to improve the performance of speech recognition system.

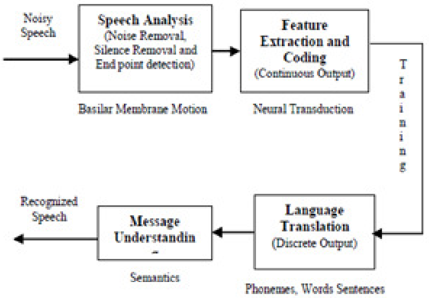
Noisy Speech
Speech Analysis (Noise Removal, Silence Removal and End point detection)
Feature Extraction and Coding (Continuous Output)
Basilar Membrane Motion
Neural Transduction
Training
Language Translation (Discrete Output)
Phonemes, Words Sentences
Message Understandin -
Semantics
Recognized Speech

Fig 1.Schematic diagram of speech recognition system

Now, in this first need to construct a signal observation representation of the test spoken utterance. The approaches in and considered a fixed length vector representation for each digit. This is done by converting the acoustic feature representation to a time-normalized representation with a fixed number of acoustic feature frames.

$$\mathbf{F}_U = (\boldsymbol{f}_{U,1}\ \boldsymbol{f}_{U,2}\ \cdots\ \boldsymbol{f}_{U,T_U})$$

2) Signal Reliability Masks:

Most works in speech processing MDT define some sort of signal reliability mask for the mel-frequency log-energy coefficients (the popularly used signal representation), which is a matrix the size of the original feature vectors with entries containing 1 to mean that the feature component is reliable and 0 to mean domination by noise.

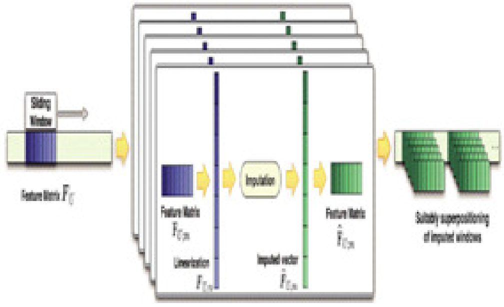

Fig 2.Diagram of sparse imputation process

VII. Possible Outcomes and Result

Initial experimentation with several window lengths showed that Tw=35 (which is also the average digit duration in the training set) is a good window length to choose for the particular database. For the frame shift parameter, it is experimented with several values using LARS-LASSO as the tuning algorithm. Note that for tuning set using LARS-LASSO Tws=1gave a recognition rate of 63.24%, Tws=5 a recognition rate of 65.04%, Tws=10 a recognition rate of 65.88% and Tws=15 a recognition rate of 45.38%. The reason behind the differences with is due to a different dictionary construction.

Speech recognition accuracy and speech recognition rate are two main terms to measure performance of speech recognition system. Speech recognition accuracy is measured in terms of Word Error Rate (WER) and speech recognition time is measured in terms of computation time. WER is a common metric of the performance of speech recognition. Here the common problem is that the recognized word sequence can have a different length from the reference word sequence.

Word error rate can then be calculated as:

$$WER = \frac{SUB + DEL + INS}{N}$$

Or

$$WER = \frac{SUB + DEL + INS}{SUB + DEL + COR}$$

For the recognition system, now use all of the8040 clean training files (containing single and continuous digit utterances) provided in the Aurora 2.0 database training set to train a continuous digit recognizer in HTK [43]. For the continuous digit recognition task, the Aurora database consists of test sets labeled N1, N2, N3, and N4 (corresponding to subway, babble, car and exhibition noise, respectively) in the Test Set A subset. By using the N1 folder for tuning of the optimization parameters. For the test sets, by creating two test sets as follows:

i) **TEST1:** merging N1, N2, N3, and N4, giving us a total of4004 files;

ii) **TEST2:** merging N2, N3, and N4 (exclusion of the N1 folder), giving us a total of 3003 files By evaluating this algorithms on different SNR conditions: SNR dB, SNR 0 dB, SNR 5 dB, and SNR 10 dB.

.

VIII. CONCLUSION

Speech is the fundamental, most effective, reliable and common medium to communicate in real time systems. There are so many applications of speech still to be far from reality just because of lack of efficient and reliable noise removal mechanism and techniques for preserving or improving the intelligibility for the speech signals.

It is shown that the smoothing effected by the decision-directed SNR estimation approach depends to a large extent on the choice and parameterization of the clean speech spectral estimator. For instance, in conjunction with the Wiener filter, hardly any smoothing is observed.

The LASSO solution for sparse imputation is relatively less effective (theoretically and experimentally) in improving the accuracies of the continuous digit recognition task as compared to the Elastic Net algorithm.

PROS

1)The speech enhancement is useful for storage and transmission of speech data, also it improves speech recognition based system performance where accurate identification of words and sentences can provide automation in most of the human-machine or machine based interface.

2) In practical applications a robust Mandarin Speech Recognition system using neural networks applied to multimedia interfaces performs better .

3)The STSA estimator that considers speech-presence uncertainty is said to produce a residual noise that is also “colorless” and does not contain musical noise.

4)It is demonstrated experimentally that by better exploiting the properties of a collinear dictionary, it is possible to expect to enjoy better speech recognition rates.

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

CONS

1)It is prove that a limitation of the a priori SNR estimate towards values in a preset range has a strong influence on both the amount of musical noise and the naturalness of the output signal.

2)The approach use a multi-style noise model trained on multiple types of noise, such an approach the performance was less than satisfactory.

APPLICATIONS

The speech enhancement is useful for storage and transmission of speech data, also it improves speech recognition based system performance. It is applied on multimedia interfaces to performs better. Musical application and hearing devices. It is used in the field of genetics and used in reconstruction of noisy speech patterns in crosstalk communication.

FUTURE SCOPE

The number of spectral segments would increase and efficient dictionary creation would become a challenge. Thus, basis selection, appropriate noise models, scarcity and algorithmic complexity will play an even more important role in large systems, and the techniques that proposed to deal with the digits task can be analogously extended to deal with a larger and more general framework.

REFERENCES

[1] Milind U. Nemade , Prof. Satish K. Shah, “Survey of Soft Computing based Speech Recognition Techniques for Speech Enhancement in Multimedia Applications”, International Journal of Advanced Research in Computer and Communication Engineering , Vol. 2, NO 5, P.P.3189 -3198, MAY 2013.

[2] Colin Breithaupt and Rainer Martin “Analysis of the Decision-Directed SNR Estimator for Speech Enhancement With Respect to Low-SNR and Transient Conditions”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, Vol. 59, No. 5, P.P. 3189 -3198, May 2011.

[3] Qun Feng Tan, Panayiotis G. Georgiou, “Enhanced Sparse Imputation Techniques for a Robust Speech Recognition Front-End”, VOL. 19, NO. 8, P.P. 2418-2430 NOVEMBER 2011.

[4] Amitai Koretz and Joseph Tabrikian "Maximum A Posteriori Probability Multiple-Pitch Tracking Using the Harmonic Model",IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 7, P.P. 2210-2221, SEPTEMBER 2011.

[5]Huijun Ding, Ing Yann Soon, and Chai Kiat Yeo "Over-Attenuated Components Regeneration for Speech Enhancement", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 8, P.P. 2004-2014, NOVEMBER 2010.

[6]Mark S. Hawley, Stuart P. Cunningham, Phil D. Green, Pam Enderby, Rebecca Palmer, Siddharth Sehgal, and Peter O’Neill "A Voice-Input Voice-Output Communication Aid for People With Severe Speech Impairment", IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, VOL. 21, NO. 1, P.P.23-31 JANUARY 2013.

[7]LIU Hong, QIAN Yanmin, LIU Jia "English Speech Recognition System on Chip", TSINGHUA SCIENCE AND TECHNOLOGY, ISSN 1007-0214 15/17 ,Volume 16, Number 1, pp95-99, February 2011.

[8]Jacob Benesty, Jingdong Chen "Binaural Noise Reduction in the Time Domain With a Stereo Setup", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 8,P.P. 2260-2272, NOVEMBER 2011.

[9]Xiaodong He,Li Deng "Speech-Centric Information Processing: An Optimization-Oriented Approach", Proceedings of the IEEE, Vol. 101, No. 5,P.P. 1116-1135, May 2013.

[10]Hynek Hermansky "Multistream Recognition of Speech: DealingWith Unknown Unknowns", Proceedings of the IEEE, Vol. 101, No. 5,P.P. 1076-1088, May 2013.