Combination of Recognizers and Fusion of Features Approach to Missing Data ASR Under Non-Stationary Noise Conditions

Neil Joshi and Ling Guan

Department of Electrical and Computer Engineering
Ryerson University
Toronto ON M5B 2K3, Canada
Email: joshi@rnet.ryerson.ca, lguan@ee.ryerson.ca

ABSTRACT

The difficulty of ASR under non-stationary noise conditions is a major contributing factor hindering the widespread deployment of ASR systems. Bottom up techniques such as speech noise separation and top down methods to adapt the acoustic model to the environment have been applied to address the issue. The missing data approach to ASR improves upon existing techniques basing recognition solely on the reliable components of the signal and has been demonstrated as an effective method to handle non-stationarity. Proposed in this paper is a novel technique whereby ASR using missing data theory under non-stationary noise conditions is improved by use of a fusion of models at the decision level. This fused model introduces more resilient features to the missing data decode process. The fused decoder is found to significantly increase recognition performance over conventional missing data techniques. A major finding in this paper is when the fused decoder exhibits the fusion of bottom up and top down processes. Under this condition, the proposed combination of recognizers technique is found to outperform all other tested ASR systems.

Index Terms—Speech recognition, Speech processing, Hidden Markov models, Pattern recognition, Time Series

1. INTRODUCTION

Over the past few decades great progress has been made in improving speech recognition with Automatic Speech Recognition, ASR, systems. In processing clean speech, systems have been designed which produce near perfect results. Even under some controlled noisy conditions, ASR systems can prove to be highly accurate. This is especially true when the corrupting variations can be considered to be stationary. Common techniques used for these conditions include Cepstral Mean Subtraction, CMN and RASTA[1]. However, a limiting factor to widespread deployment of ASR based systems is in handling the majority of the noise conditions, when the corrupting factors are non-stationary. Recently, ASR using Missing Data, MD, techniques has been proposed as a method for noise robust ASR under all noise conditions.

The use of missing data techniques in ASR is based on the premise that recognition should only be conducted upon speech bearing components of a signal. The underlying principal relies on evidence on how the human auditory system is believed to perceive and process speech. Here all composite signals bearing both speech and noise are segregated and processed individually. Spectral features are used with standard missing data ASR, though, spectral based features for HMM based ASR systems are known to be deficient in resiliency to slight perturbations. Theses features are used within MD systems due to its representation maintaining a level of detail necessary for auditory source separation, the central premise of missing data theory. Mel-frequency Cepstral Coefficient, MFCC, features have been established to be well suited for HMM based ASR. Inherent to the mel-frequency transformation process statistical variations are removed, thus allowing a more robust HMM representation. Unfortunately, the use of MFCC based features in missing data theory based ASR systems has not been suitable due to "smearing" of localized uncertainties globally in the auditory signal. Thus the very characteristic that allows MFCC features to be well suited for HMM based systems hinders their use in missing data theory.

This paper proposes a method to enhance speech recognition performance using missing data techniques for non-stationary noise conditions by incorporating more resilient feature sets into the decoding process. Within this process, the effects of the introduction of resilient features and noise compensation techniques to the MD ASR decode process are realized. This is accomplished by the creation of two separate HMM based models, one using spectral features, the other MFCC features. The statistical dependencies found in the models are based upon a coupled HMM methodology, the Fused HMM model[2]. This fusion of features and combination of recognizers, one using standard ASR techniques, the other missing data based, is demonstrated in this work to significantly increase recognition performance when speech bearing signals are corrupted by non-stationary additive noise. This is particularly evident when the fused decoder exploits the fusion of a noise compensated acoustical model and the standard MD model. Under this condition, the proposed combination of recognizers technique is found to outperform all other tested ASR systems.

There exist a plethora of research into enhancing ASR with missing data theory. The majority of the approaches have concentrated on improving the separation of speech from noise prior to the decode process. Soft masks[3] have been developed to assign probabilities to each component within the source segregation mask prior to decoding. More sophisticated methods have been applied to aid in segregation of speech from noise such as employing auditory scene analysis techniques[4]. With regard to improvements to the MD decoding process, the Multisource decoder[5] has been proposed where the auditory signal is broken into fragments and the best hypothesis is based upon finding the best word match and the best segregation of speech from noise. The feature fusion technique has been applied to improve recognition performance using both spectral and MFCC based features under Stationary additive noise conditions[6]. This paper does not address improving segregation of signals to increase recognition performance, rather it proposes a method to enhance recognition performance by utilizing statistical dependencies between MD theory spectral HMM models and traditional MFCC based models to produce an improved fused acoustical model. This exploitation of complementarity[7] information between the two models is demonstrated to significantly increase recognition performance.

The structure of this paper is organized in the following manner. First, approaches to enhance ASR recognition performance under

non-stationary noise conditions is presented outlining existing methods and the approach presented in this paper. Next the proposed feature fusion, combination of recognizers, method is detailed and described as how to be applied to increase recognition performance with missing data techniques. A series of experiments that were conducted to demonstrate the application of the proposed method is described, followed by the results from those experiments. Finally, conclusions and directions for further research are discussed.

2. FUSION OF FEATURES AND COMBINATION OF RECOGNIZERS

2.1. Non-Stationary Noise Robust ASR

Established methods for achieving noise robust ASR involving noise/speech separation generally take on the form of one of two methodologies. Ones that attempt to compensate for distortions found in signals by the extraction of the clean signal prior to recognition by the ASR decoder can be regarded as a bottom-up approach. The second method of designing a system for handling non-stationary noise perturbations is developed from a top down model basing noise robust ASR as a pattern recognition problem. The bottom-up approach is designed to segregate the noise from the speech to match with the acoustical models of the ASR system. The top-down approach attempts to produce a matched system. A matched system has acoustical data or models which are identical to that of the acoustical characteristics of the input sequence. If the recognition task is such that the input conditions are constant, then the acoustical models may be trained in that environment. Under variable input acoustical conditions, more realistic, a matched system may be achieved through the use of techniques which base recognition on the acoustical model transformed by the given noise source. Such methods which transform acoustical models are the HMM Decomposition[8] and Parallel Model Combination[9], PMC methods.

With regard to ASR using missing data techniques for non-stationary noise conditions, both bottom-up and top-down approaches have been investigated. Bottom-up approaches consist of generating a segregated noise mask employing spectral subtraction, soft-mask generation, and auditory scene analysis. Recently, a top-down noise compensation technique, equivalent to the highly successful CMN normalization in the cepstral domain, has been proposed for convolutional noise[10]. The multisource decoder, developed specifically to enhance recognition performance under non-stationary conditions, utilizes both bottom-up and top-down methods to find the best hypothesis word sequence and segregation. The proposed method in this paper approaches increasing recognition accuracy using MD techniques in a novel perspective. In this procedure, the resiliency of the model used in the decode process is elevated. This is achieved by means of using additional feature sets which add complementarity information to the viterbi search space in determining the best hypothesis word sequence.

2.2. Proposed Methodology

ASR with missing data theory consists of three major components, namely, feature extraction, segregation and the decoder. HMM models created with use of spectral based features, though, cannot be fully representative of the auditory characteristics when modeled by the typical diagonal covariance matrix. Even with the addition of multiple mixtures per model, the robustness of such a model does not achieve the accuracy of its MFCC based counterpart. Thus there is a case for incorporating the use of these resilient features in the MD ASR process. The proposed method addresses this issue by presenting a procedure whereby the benefits of ASR with missing data techniques can be combined with the robustness of standard ASR systems. This is accomplished with use of the combination of recognizers. The combination of recognizers is formed on the premise of the fusion of features at the decision level. Decision level fusion permits the contributions of each separate stream of features to be realized in the decode process. The chosen method of decision level fusion is conducted as the stochastic modeling of coupled time series. With the combination of recognizers approach MFCC derived features from an auditory signal can be incorporated into a missing data theory based ASR system using a coupled HMM methodology. For this model, several different coupled HMM models were considered. The most accommodating was found to be the Fused HMM model[2].

The fused HMM model models the relationship between HMMs using a probabilistic fusion model. With this method, optimal HMM connections are made using the maximum entropy principal and maximum mutual criterion for selecting dimension reduction transforms. Using the maximum entropy principal, the joint distribution between observations of two time series, $\mathbf{O^{(1)}}$ and $\mathbf{O^{(2)}}$ and hidden states $\mathbf{U^{(1)}}$ and $\mathbf{U^{(2)}}$ respectively is defined to be,

$$\hat{p}(O^{(1)};O^{(2)}) = p(O^{(1)})\,p(O^{(2)})\frac{p(w,v)}{p(w)\,p(v)} \tag{1}$$

The optimal HMM connections are found by the determination of **w** and **v** with the application of the maximum mutual information criterion. Here it is found that with the relation,

$$I(f(x),y) \le I(x,y) \tag{2}$$

the optimal HMM connections are between the observations of one HMM model and the hidden states of the other. The resulting expression describing the statistical dependencies between two HMM processes is thus,

$$\hat{p^{(1)}}(O^{(1)};O^{(2)}) = p(O^{(1)})\,p(O^{(2)}|\hat{U^{(1)}}) \tag{3}$$

and,

$$\hat{p^{(2)}}(O^{(1)};O^{(2)}) = p(O^{(2)})\,p(O^{(1)}|\hat{U^{(2)}}) \tag{4}$$

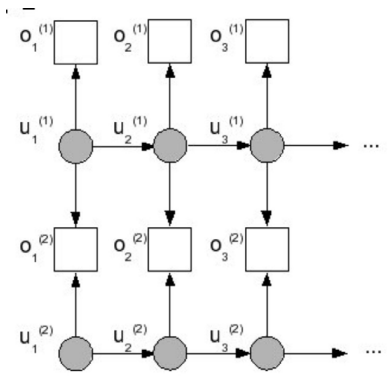

Fig. 1: Coupled Fused HMM

The resulting Fused HMM model composed of two HMM processes is depicted in Fig. 1 illustrating the interconnections between the two models.

Using the fused HMM model, the combination of recognizers technique permits MFCC derived features from an auditory signal to be incorporated into a missing data theory based ASR system. The acoustic model created with MFCC derived features and the model created with traditional missing data theory techniques, spectral features, are fused together to create an optimal model. Recognition is then performed by the combination of recognizers by generating

both spectral and MFCC features from an input auditory signal and decoded with the fused HMM decoder.

3. EXPERIMENTS

3.1. Baseline Recognizer

A baseline recognizer was created with the Grid Corpus[11]. All experiments conducted in this paper are formed from the creation of a HMM based recognizer consisting of 51 words as depicted in Fig. 2.

command	color	preposition	letter	number	adverb
bin (b) lay (l) place (p) set (s)	blue (b) green (g) red (r) white (w)	at (a) by (b) in (i) with (w)	A – Z excluding W	1-9 and zero (z)	again (a) now (n) please (p) soon (s)

Fig. 2: Corpus Vocabulary

Word level CDHMM models were used consisting of 2 states per phoneme in accordance with the CMU pronunciation dictionary. Each state within the model was composed of 32 Gaussian Mixtures. Training of the recognizer was conducted using 17000 unique sentences from 17 different speakers with 1000 sentences generated from each speaker. Within the corpus, each utterance was composed of 6 words with the grammar corresponding to Fig. 3.

```
$verb=bin|lay|place|set;
$colour=blue|green|red|white;
$prep=at|by|in|with;
$letter=a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|x|y|z;
$number=zero|one|two|three|four|five|six|seven|eight|nine;
$coda=again|now|please|soon;

($verb sp $colour sp $prep sp $letter sp $number sp $coda)
```

Fig. 3: Corpus Utterance Grammer

Three acoustic models were constructed using the training corpus. The first, a spectral ratemap based model. The others cepstral domain, MFCC based models. Feature vectors used for the spectral model consist of ratemaps produced by passing the auditory signal through a bank of 32 Gammatone Filters with center frequencies spaced linearly in ERB-rate from 50Hz to 3850Hz. The envelope of the output from each filter was smoothed with an 8ms time constant and sampled at a frame rate of 10ms. The cepstral based models used were composed of 39 dimensions consisting of energy, delta and acceleration coefficients. All CDHMM acoustic models were constructed using the HTK Toolkit[12].

3.2. Experiments Setup

A speech recognizer was setup to analyze recognition results from a testing corpus using spectral, cepstral and the proposed combination of recognizers, COR, fusion, based features. The testing corpus consisted of 560 utterances generated by 14 different speakers of which 40 contributed from each speaker. The speakers used in the testing corpus are independent to that used in the training set. The recognizer was setup in 8 different configurations to perform ASR with,

I. spectral, ratemap features, rate32
II. cepstral features, MFCC
III. cepstral features with normalization, CMN
IV. ratemap using missing data techniques, MD
V. COR, rate32+MFCC
VI. COR, rate32+CMN
VII. COR, MD+MFCC
VIII. COR, MD+CMN

The experiments to determine the gain in speech recognition performance using the combination of recognizers and feature fusion proposed method over conventional MD techniques for non-stationary noise conditions were conducted in the following manner. The testing corpus was corrupted with additive non-stationary noise with various SNRs to form test sets. The noise source was taken from the NOISEX[13] database and consisted of the Factory Noise I source. The noise source was added to the testing corpus to form 3 separate test sets of SNRs being 18dB, 12dB and 6dB. The choice of the noise source was to demonstrate the performance of the proposed method for an extreme example of non-stationarity noise.

3.3. Results

The results from running the recognizer in configurations I to III and V to VI with the clean speech test corpus is illustrated in Table I. A thorough analysis of the fused decoder results with clean speech is addressed in a prior paper[6].

TABLE I. BASELINE RECOGNIZER RESULTS, CLEAN DATA

ASR Configuration	Recognition Accuracy, %
MFCC	95.15
MFCC CMN	81.49
Spectral Features, rate32	94.64
MFCC+rate32	94.04
MFCC CMN+rate32	95.22

Under non-stationary noise conditions, Table II depicts recognition results with the recognizer in configurations I to IV and VII to VIII with the test corpus corrupted with additive Factory noise respectively at all tested SNRs. For higher SNRs the recognizer

TABLE II. RECOGNITION RESULTS WITH TEST CORPUS + FACTORY NOISE

		SNR 18 dB	SNR 12 dB	SNR 6 dB
Conventional	MFCC	83.33	73.9	64.7
	MFCC CMN	66.0	61.6	60.3
	Spectral Features, MD	76.5	73.3	67.4
Proposed Method	COR, MFCC+MD	84.5	76.7	67.5
	COR, MFCC CMN+MD	88.6	81.8	73.5

configured with MFCC features outperforms the recognition accuracy achieved using classical MD techniques. As the SNR is decreased, the benefits of ASR with MD is realized as found with the 6dB conditions, the recognizer using MD techniques surpasses accuracies derived from MFCC features. The acoustic model for the MFCC based recognizer is not able to properly describe all of the auditory input conditions for lower SNRs and thus recognition accuracies decline more rapidly than that of the recognizer using MD theory. Using the COR technique to base recognition upon both MFCC and spectral features, the accuracies achieved exceed or match those by just MD alone demonstrating the validity of enhancement by use of complementarity information in the decode process. As the SNR decreases the MFCC acoustic model no longer accommodates the characteristics inherent to the input signal, thus there exists less improvement with the COR technique. In this case, the MFCC acoustic model contributes less complementarity and more supplementarity[7] information to the decode process. In contrast, recognition using noise robust techniques such as top down modeling, to adapt the acoustical model for the input auditory conditions, in this case normalization, greatly enhances recognition performance in conjunction with MD techniques. This is evident over all tested noise conditions. In this mode, both bottom up and top down techniques are used to achieve noise robust ASR under non-stationary noise conditions. Bottom up in terms of segregation of noise from speech prior to decoding only the reliable components in the signal. Top down employed by means of normalization using the global mean of the signal to compensate the model to the variable acoustical conditions.

Tables III to V depict the performance of the recognizers rankings on a per utterance basis for configurations of interest with each of the three test sets. The rankings are based upon the

recognition accuracy obtained by the ASR of each utterance with the recognizer in configurations II to IV and VII to VIII. From examination of each recognizer performance, one can deduce that

TABLE III. RECOGNIZER CONFIGURATION RANKINGS OVER ENTIRE TEST SET, 18DB SNR, RELATIVE TO ALL EXPERIMENTED CONFIGURATIONS

ASR Configuration	# utterances Top Ranked	% Top Ranked	#utterances Bottom Ranked	%Bottom Ranked
Missing Data	109	19.64	161	29.01
COR MFCC	230	41.44	30	5.41
COR CMN	390	70.27	16	2.88

TABLE IV. RECOGNIZER CONFIGURATION RANKINGS OVER ENTIRE TEST SET, 12DB SNR, RELATIVE TO ALL EXPERIMENTED CONFIGURATIONS

ASR Configuration	# utterances Top Ranked	% Top Ranked	#utterances Bottom Ranked	%Bottom Ranked
Missing Data	142	25.59	139	25.05
COR MFCC	193	34.77	60	10.81
COR CMN	383	69.01	29	5.23

TABLE V. RECOGNIZER CONFIGURATION RANKINGS OVER ENTIRE TEST SET, 6DB SNR, RELATIVE TO ALL EXPERIMENTED CONFIGURATIONS

ASR Configuration	# utterances Top Ranked	% Top Ranked	#utterances Bottom Ranked	%Bottom Ranked
Missing Data	189	34.05	181	32.61
COR MFCC	161	29.01	138	24.86
COR CMN	368	66.31	54	9.73

using the proposed method does not degrade the performance established by standard MD techniques. The number of utterances that are ranked the lowest when compared to all tested recognizer configurations shows the fusion technique to consistently be less than that of processing with MD. The COR technique has a profound effect on influencing the increase in recognition performance when decoding with the fused information from noise compensated cepstral features and MD techniques. This is evident upon examination of the number of utterances that have been recognized the most accurate by each recognizer.

Here it must be stated that the results obtained using the proposed technique addresses concerns raised in previous attempts to enhance the decoding by means of Data Imputation[14]. One of the findings was that normalization in conjunction with MD using marginalization degraded recognition performance. Demonstrated in the results presented with the combination of recognizers technique, ASR using MD with marginalization is significantly enhanced when fused with a noise compensated model.

The combination of recognizer technique in the configuration of the fusion of the MD decode process with complementarity information from a normalized cepstral model is found to significantly increase ASR performance. This configuration has been presented to be akin to the combining of a bottom up process and top down process to address the non-stationary noise condition. Currently there exists one other approach that advances ASR using MD techniques using a combination of these two processes, the fragment decoder[5]. Whereas the multisource decoder finds the best hypothesis word sequence and segregation of noise from speech, the combination of recognizers finds the best word sequence using a fused decoder. This fused decoder exploits the statistical dependencies between an adapted acoustic model and an MD marginalized decode process. Like the multisource decoder exhibiting significant gains in recognition performance when subjected to non-stationary noise conditions, the COR with CMN and MD performs especially well under these conditions. This is clearly illustrated by comparing the results presented in this paper with COR under stationary noise conditions[6].

4. CONCLUSIONS

Recognition accuracy of ASR using missing data techniques under non-stationary noise conditions is improved using the proposed combination of recognizers technique. Recognition is improved or matched with the use of resilient MFCC based feature sets fused with the MD decode process. Under conditions where a noise compensated model is used in conjunction with standard MD techniques significant performance gains are realized. Specifically, as outlined in the paper, this occurs when the combination of recognizers is configured using CMN within the fused decoder. This configuration establishes the use of both bottom up and top down ASR techniques and proves to be highly successful. Future work will entail improving upon speech enhancement of the input signal by means of using more sophisticated segregation techniques. This shall improve upon the missing data decode process within the combination of recognizers technique. Investigations will also conducted in incorporating additional noise compensation techniques to build upon the results presented in this paper.

5. REFERENCES

[1] H. Hemansky, "RASTA Processing of Speech", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-594, 1994.

[2] H. Pan, S. Levinson, T. Huang, Z. Liang, "A Fused Hidden Markov Model With Application to Bimodal Speech Processing", *IEEE Transactions on Signal Processing*, Vol. 52, No. 3, March 2004.

[3] J. Barker, L. Josifovski, M.P. Cooke and P.D. Green, "Soft Decisions in Missing Data Techniques for Robust Automatic Speech Recognition", *Proceedings ICSLP*, 2000.

[4] G.J. Brown, M.P. Cooke, "Computational Auditory Scene Analysis", *Computer Speech and Language*, vol. 8, pp. 297-336, 1994.

[5] J. Barker, M.P. Cooke, D. Ellis, "Decoding Speech in the Presence of Other Sources", *Speech Communications*, No. 45, pp. 5-25, 2005.

[6] N. Joshi, L. Guan, "Missing Data ASR with Fusion of Features and Combination of Recognizers", *to appear in Proceedings of Workshop on SLT*, 2006.

[7] C. Chilbelushi, F. Deravi, "A Review of Speech-Based Bimodal Recognition", *IEEE Trans. On Multimedia*, vol. 4, no. 1, March 2002.

[8] A.P. Varga, R.K. Moore, "Hidden Markov Decomposition of Speech and Noise", *Proceedings ICASSP*, pp. 845-848, 1990.

[9] M. Gales, S. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination", *IEEE Trans. Speech and Audio Processing*, pp. 352-359, 1996.

[10] K.J. Palomaki, G.J. Brown, and J. Barker, "Techniques for Handling Convolutional Distortion with Missing Data Automatic Speech Recognition", *Speech Communications*, vol. 43, no. 1-2, pp. 123-142, 2004.

[11] M. P. Cooke, J. Barker, S.P. Cunningham, and X. Shao, "An Audio-Visual Corpus For Speech Perception and Automatic Speech Recognition", *submitted to JASA.*

[12] S. Young, P. Woodland, "HTK Version : User, Reference and Programmer Manual," Cambridge University Engineering Department, Speech Group, 2002.

[13] A. P. Varga, H. J.M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition", *Tech. Rep., DRA Speech Res. Unit*, 1992.

[14] B. Raj, M.L. Seltzer, and R.M. Stern, "Robust Speech Recognition: The Case For Restoring Missing Features", *Proceedings Eurospeech, The Workshop on Consistent and Reliable Acoustic Cues*, 2001.