Missing Data ASR with Fusion of Features and Combination of Recognizers

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

Speech recognition under noisy conditions has been actively researched and effective techniques have been developed to handle stationary noise. Under circumstances where the stationary assumption is not valid, the performance of speech recognizers is extremely poor. Missing data theory provides a method for the development of robust speech recognition under any noisy condition. A limitation to ASR with missing data theory techniques is the choice of features used in the model. There exist alternative feature representations that have been demonstrated to be much more effective for signal recognition purposes. This paper presents a novel method to incorporate the use of alternative feature sets within the realm of ASR with missing data theory techniques. Using the proposed combination of recognizers, or fusion of features, an ASR decoding process is developed based upon the coupling of spectral features using missing data techniques and traditional MFCC based features. The proposed technique is demonstrated to increase recognition performance under all experimented noise conditions over traditional missing data techniques.

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

1. INTRODUCTION

The performance of systems for Automatic Speech Recognition Systems, ASRs, have been greatly advanced over the past few decades. ASRs though, still perform poorly when processing speech signals that have been distorted by noise. Great progress has been made in developing robust ASRs for situations where interfering noise may be considered to be Stationary. Advanced speech processing techniques such as Cepstral Mean Normalization, CMN, RASTA, and spectral subtraction have been developed to improve recognition performance. These widely used methods though, have been shown to only improve performance when the noise disturbance to the speech signal can be regarded to be stationary. Unfortunately to realize truly noise robust ASR, a method is required to be developed to compensate for speech signals distorted by the majority of noise disturbances, the non-stationary case. The missing data approach to achieving noise robust ASR is one such method. ASR with missing data techniques has been demonstrated to provide improved ASR results over conventional recognizers under most conditions, though has not performed as well with higher SNRs and for stationary noise conditions[1].

The benefits of the ASR with missing data techniques includes that it provides comparable recognition to traditional techniques under most conditions and superior accuracy in non-stationary noise conditions. Unlike methods developed to handle non-stationary noise conditions, such as HMM Decomposition and Parallel Mode Combination, PMC, this method makes no assumptions regarding the type of disturbance found in auditory signals nor does it require any noise models present in the recognizer. Regardless of the distortion found in the signal, missing data techniques segregates speech from the noise and recognition is performed using the segregated speech.

Limitations to ASR with missing data techniques can be attributed to the feature set used in the recognition system. Auditory spectral features are used as they provide a representation that parallels the method employed by the human auditory system in the perception of speech. These features are suitable for auditory source separation which is vital to missing data techniques. Spectral features though are known to be not as resilient as features derived in the cepstral domain[2]. In the cepstral transformation process variations between features in the feature set are removed. This statistical independence allows HMM models with diagonal covariance to be a suitable model for recognition. The statistical independence assumption between features does not hold as strongly for spectral features, thus attributes to decreased recognition performance. Preprocessing techniques such as the normalization of features, has been extremely successful in improving ASR accuracy when applied to cepstral based features. The normalization process though, is not compatible with spectral based features, though a technique has been recently proposed to address this issue[3].

This paper presents a method whereby the benefits of ASR with missing data theory techniques can be incorporated into a new ASR model exploiting complementarity[4] information found in traditional ASRs. This new model that is a *combination of recognizers* or fusion of existing models is demonstrated through a series of experiments to improve recognition under noisy conditions. This new model, for recognition, is based upon the stochastic modeling of coupled time-series in terms of HMMs formed from cepstral and spectral based features. Specifically, the fusion of a mel-frequency cepstral coefficient, MFCC, based recognizer and a recognizer based upon missing data techniques. The formation of this new model is accomplished using the Fused HMM method[5].

This paper is presented in the following format. The proposed combination of recognizers is described. The experimental setup used for comparison of recognition results is then described followed by the results of ASR with the combination of recognizers technique. Finally, a discussion of issues and topics for future work is described.

2. COMBINATION OF RECOGNIZERS

Noise robust ASR can be effectively achieved through the use of missing data theory techniques. The central premise is to separate the acoustical mixture into the desired signal and noise. Recognition accuracy may be improved upon over existing methods by the use of alternative superior feature representations of the auditory signal or by the fusion of features as proposed in this paper.

2.1. MFCC Based Features - Desirable Features

Features used in traditional missing data theory based ASRs are based upon spectral representations of the auditory signals. Specifically ratemap, time frequency intensity mappings of a signal depicting the level of neurons firing in the cochlear. It is well established that such features, though extremely effective in auditory scene analysis, are not as resilient as the use of Mel-Frequency Cepstral Coefficient, MFCC, derived features for recognition purposes. Reasons for this are derived from the suitability of MFCCs in HMM based recognizers using diagonal covariance matrix representations. Within the MF transformation process statistical variations between features are removed allowing this feature representation to be accurately modeled in this manner. Spectral features, though, require a full covariance HMM representation to be accurately modeled. To compensate for this and to allow the use of diagonal covariance HMM models with spectral features, it has been proposed to use multiple Gaussian Mixtures within each HMM state. The use of MFCC based features directly in missing data theory based ASR systems has not been suitable due to "smearing" of localized uncertainties in the auditory signal globally in the MF transformation process[1]. The identification of localized uncertainties is crucial in missing data theory which relies on the determination of reliable and unreliable components within an auditory signal.

2.2. Existing Techniques in using MFCC features with Missing Data

The use of the well established MFCC based features is highly desirable in noise robust ASR systems. Currently there exist two methods to address the use of MFCC derived features in ASR with missing data techniques,

1. direct method
2. indirect method

The first, the direct method, applies missing data techniques to ASR by operating in the cepstral domain. The segregation of noise and speech translates to the application of cepstral distance weights[6] for each of the Mel-frequency subbands of the transformed signal. The decode process uses marginalization techniques[7]. Results from this method have proved be unsuccessful for noise robust ASR. The second method, employs the data imputation technique in decoding with missing data theory[7]. Here, the segregation of the signal is conducted in the spectral domain. An estimated restored signal is then transformed into the cepstral domain prior to decoding. Results from this method have been shown to be less accurate than recognition using traditional missing data techniques.

The proposed combination of recognizers technique, is based upon the stochastic modeling of coupled time-series in terms of HMMs formed from cepstral and spectral based features. Specifically, the fusion of a mel-frequency cepstral coefficient, MFCC, based recognizer and a recognizer based upon missing data techniques. The statistical relationship between the missing data spectral model and a cepstral domain model is determined by coupling observations in one domain to the hidden states in the other.

2.3. The Combination of Recognizers Methodology

The use of separate feature sets in recognition of speech signals has been vigorously researched recently. Popular incarnations of these recognizers use features extracted from audio and synchronized video frames utilizing information from these two modalities to create a *bimodal recognizer*. Borrowing concepts from bimodal recognizer systems, the use of MFCC derived features from an auditory signal can be incorporated into a traditional missing data theory based ASR system. This combination of recognizers contains two streams of features. With this new model there exists both supplementarity and complementarity information in relation to the separate feature sets. Complementarity in terms of one of the two sets containing additional information pertaining to the auditory signal that the other lacks. Supplementarity information as there exist inherent redundancies in the content.

Modeling the contributions of multiple features in a recognition system can be conducted in various manners. The integration of features may be viewed as low-level integration or intermediate/late integration. Low-level integration involves feature fusion where a feature vector is formed based upon the concatenation of features from the different streams. The dimensionality of this new feature space is significantly greater than the originating feature sets. For this reason PCA or LDA is conducted to reduce the dimensionality. In intermediate/late integration, the fusion is referred to as *decision fusion* where a model is formed based upon the dependencies between the models of the individual streams. In terms of combining the missing data theory based recognizer and a MFCC based recognizer, a decision fusion technique is chosen as this is the only method which may accommodate the use of missing data theory decoding by means of marginalization. With this premise, the family of decision fusion techniques involving the stochastic modeling of coupled time series is appropriate.

Using HMM based methods, a model may be constructed as a signal HMM or the coupling of multiple HMMs by the use of Coupled HMM[8], Mixed-Memory HMM[9], or the recently proposed Fused HMM[5] manners. A single HMM model is inappropriate for the fusion of features due to incorporating multiple sets of features into as single model may not be practical and suffers form the overfitting problem. Within the family of coupled HMMs, the basic structure exhibits the statistical dependencies between models by linking the hidden states of each model. The Mixed-Memory HMM enhances the dependencies by linking the hidden states in one HMM to the observation of the other. The fused HMM model models the relationship between HMMs by using a probabilistic fusion model. With this method, optimal HMM connections are made using the maximum entropy principal and the maximum mutual criterion for selecting dimension reduction transforms. The fused HMM method is chosen for the combination of recognizers as it determines an optimal coupled representation. It lends itself well to conventional HMM based ASR decoders as it uses well established techniques for training, EM, and decoding, viterbi. Also the technique is computationally efficient and robust as dependencies between HMMs are determined from separately trained HMMs. Should one of the models prove to be unstable, the coupled HMM fusion model is still functional as information is derived from other HMMs in the system.

The fused HMM model is illustrated in the following derivation. Let $\mathbf{O^{(i)}}$ be one of **i** time series consisting of ($\mathbf{o^{(i)}_0}$, $\mathbf{o^{(i)}_1}$, $\mathbf{o^{(i)}_2}$, ...) and modeled as HMMs with hidden states, $\mathbf{U^{(i)}}$. Define two HMMs, **1** and **2** by,

$$p(O^{(1)}, U^{(1)}) = p(u_0^{(1)}) \prod_{t=1}^{T-1} p(u_t^{(1)} | u_{t-1}^{(1)}) \prod_{t=1}^{T-1} p(o_t^{(1)} | u_t^{(1)}) \quad (1)$$

$$p(O^{(2)}, U^{(2)}) = p(u_0^{(2)}) \prod_{t=1}^{T-1} p(u_t^{(2)} | u_{t-1}^{(2)}) \prod_{t=1}^{T-1} p(o_t^{(2)} | u_t^{(2)}) \quad (2)$$

The statistical dependence between the two HMMs, the joint distribution of the system, is found using two transforms, **w** and **v** and is described by the following relationship,

$$\hat{p}(O^{(1)}; O^{(2)}) = p(O^{(1)}) p(O^{(2)}) p \frac{(w, v)}{p(w) p(v)} \quad (3)$$

which can be further expressed as,

$$\hat{p}(O^{(1)};O^{(2)}) = 0.5\,\hat{p^{(1)}}(O^{(1)};O^{(2)}) + 0.5\,\hat{p^{(2)}}(O^{(1)};O^{(2)}) \quad (4)$$

where,

$$\hat{p^{(1)}}(O^{(1)};O^{(2)}) = p(O^{(1)})\,p(O^{(2)}|\hat{U^{(1)}}) \quad (5)$$

and,

$$\hat{p^{(2)}}(O^{(1)};O^{(2)}) = p(O^{(2)})\,p(O^{(1)}|\hat{U^{(2)}}) \quad (6)$$

Within this model, for a given time, the likelihood of a hidden state generating each of the two streams of observations is evaluated as illustrated in Fig. 1.

MFCC derived features from an auditory signal can be incorporated into a missing data theory based ASR system using a coupled HMM methodology. 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 a coupled model. Using the HMM fusion method, an optimal model exhibiting the statistical dependency between the two models are created. Recognition is then performed by creating both spectral and MFCC features and decoded with the Fused HMM decoder. The new ASR system using missing data theory in combination with MFCC features is depicted in block diagram form in Fig. 2.

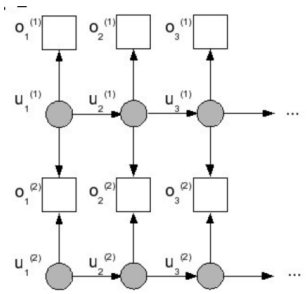

Fig. 1: Coupled Fused HMM

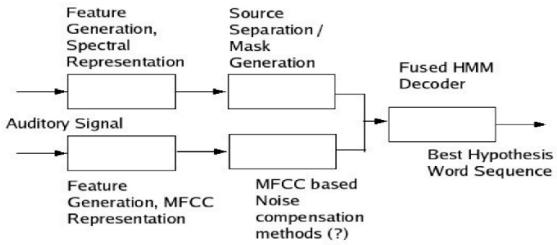

Fig. 2: Proposed fused HMM model

3. EXPERIMENTS

A baseline recognizer was created for experimental result comparisons. The Grid Corpus[10] was used for training and testing the recognizer. All experiments conducted in this paper are formed from the creation of a HMM based recognizer consisting of 51 words. Each word is modeled as a CDHMM consisting of 2 states per phoneme. Each state within the model consists of 32 Gaussian Mixtures. Training of the recognizer consists of 17000 unique sentences from 34 different speakers with 500 sentences generated from each speaker.

3.1.Experiments Setup

Three acoustical models were constructed using the training corpus, a spectral and two MFFC based. Feature vectors used for the spectral model consisted 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 MFCC features that were 39 dimensional consisting of energy, delta and acceleration coefficients. All CDHMM acoustical models were constructed using the HTK Toolkit[11].

A speech recognizer was setup to analyze recognition results from a testing corpus using spectral, cepstral and the proposed combination of recognizers, *fusion*, based features. The testing corpus consisted of 300 utterances from 34 different speakers. Each of the 300 utterances tested are not contained within the training corpus. The recognizer was setup in 8 different configurations to perform ASR with,

i. spectral, ratemap features, rate32
ii. cepstral features, MFCC
iii. cepstral features normalized, CMN
iv. ratemap using missing data techniques, MD
v. fusion, rate32+MFCC
vi. fusion, rate32+CMN
vii. fusion, MD+MFCC
viii. fusion, MD+CMN

Recognition was conducted with configurations i.-vi. for the test set and for configurations i.-iv. and vii.-viii. for the test set when subjected to stationary noise with a SNR of 6dB and 0dB.

The recognizer used in all experiments was a custom recognizer derived from the CTK Toolkit[12]. For ASR using missing data techniques, segregation of speech from noise was conducted based upon a noise floor from a noise estimate derived from the first few frames of the utterance processed.

3.2.Results

Results from experiments conducted to validate the premise of increased speech recognition performance with feature fusion over current missing data techniques are outlined in this subsection. Using the experimental setup described in the previous subsection, Table I. depicts the recognition results for all ASR configurations under all noise conditions. Recognition accuracy is based upon the percentage of correct words recognized over each utterance in the corpus. For the recognition using a clean test corpus, no noise condition, fusion was conducted with ratemap features and cepstral features. For recognition under all other conditions of the test corpus, fusion was conducted using a standard spectral missing data based ASR and cepstral features.

TABLE I. RECOGNITION % ACCURACY, RESULTS COMPARING EXPERIMENTED RECOGNIZER CONFIGURATIONS UNDER VARYING NOISE CONDITIONS

		SNR Stationary Noise		
ASR Configuration		6dB	0dB	clean
Conventional	Spectral Features, rate32	66.9	60.9	96.6
	MFCC Features	78.2	64.9	98
	CMN	69.4	62.1	95.6
	ASR with Missing Data, MD	76.4	69.8	N/A
Fusion, Combination of Recognizers	MFCC+rate32	N/A	N/A	97.8
	CMN+rate32	N/A	N/A	97.8
	MFCC+MD	82.4	71.3	N/A
	CMN+MD	82.7	72.2	N/A

From Table I, it is clearly evident that fusion of features outperforms recognition using spectral based features and all other configurations. For high SNRs, 6dB, a substantial increase is demonstrated using fusion of both missing data, MD and cepstral based methods. It is demonstrated that supplementary information contained in one feature set is enhancing recognition performance when stochastically coupled with the other. In this case the coupling is between the observations in one domain tied with the likelihood of being generated by the hidden states in the other domain.

Reasons for the small margin of increased recognition performance using the fusion technique over others for the clean test corpus can be attributed to the already high level of accuracy achieved by all ASR configurations and thus the smaller amount of supplementary information in one feature set over the others. The level of increased recognition performance of the the combination of recognizers method over all ASR configuration for the 0dB case may be due to the recognition performance of ASR with standard missing data techniques. Here, under this condition, the degree of accuracy is low and is due to the method employed to segregate the speech from noise. Currently, a noise estimate is conducted by examining the first few frames of each utterance and segregation is made by using a noise floor with this estimate. The test corpus used is end pointed, therefore containing little or no silence prior to the commencement of speech. The noise estimate thereby is crude and most likely not suitable for this condition. It is anticipated that a substantial increase in recognition performance can be achieved using more sophisticated methods to obtain a noise estimate and in turn a much higher recognition increase with the fusion method.

To further analyze the increase in recognition performance achieved by the combination of recognizers technique over traditional ASR with missing data techniques, rankings based upon recognition accuracy for each utterance in the test corpus were conducted for the 6 different configurations. Out of the 300 utterance corpus, the percentage that a particular ASR configuration is ranked the highest and lowest for each noise condition is illustrated in Tables II - IV for MD and fusion methods.

TABLE II. RECOGNIZER CONFIGURATION RANKING OVER ENTIRE TEST CORPUS RELATIVE TO ALL EXPERIMENTED ASR CONFIGURATIONS, 6DB NOISE CONDITION

Noise Condition, 6dB				
ASR Configuration	# utterances Top Ranked	% Top Ranked	# utterances Bottom Ranked	% Bottom Ranked
Missing Data	90	30	57	19
Fused MFCC	180	60	17	5.67
Fused CMN	170	56.67	8	2.67

TABLE III. RECOGNIZER CONFIGURATION RANKING OVER ENTIRE TEST CORPUS RELATIVE TO ALL EXPERIMENTED ASR CONFIGURATIONS, 0DB NOISE CONDITION

Noise Condition, 0dB				
ASR Configuration	# utterances Top Ranked	% Top Ranked	# utterances Bottom Ranked	% Bottom Ranked
Missing Data	126	42	49	16.33
Fused MFCC	135	45	28	9.33
Fused CMN	158	52.67	33	11

TABLE IV. RECOGNIZER CONFIGURATION RANKING OVER ENTIRE TEST CORPUS RELATIVE TO ALL EXPERIMENTED ASR CONFIGURATIONS, NO NOISE CONDITION

Noise Condition, Clean				
ASR Configuration	# utterances Top Ranked	% Top Ranked	# utterances Bottom Ranked	% Bottom Ranked
Spectral, rate32	90	30	64	21.33
Fused MFCC	240	80	47	15.67
Fused CMN	241	80.33	47	15.67

Clearly illustrated from the rankings of the occurrence from the test corpus that a particular configuration achieves the highest, top ranked, or lowest, bottom ranked, recognition accuracy one can extrapolate that the fusion technique is enhancing recognition performance. The low percentage of the bottom rank for the fusion technique shows that fusion will at the very least maintain the current level of accuracy achieved using traditional means.

4. CONCLUSIONS

Missing data theory techniques enhances recognition under noisy conditions but employs non ideal features in its ASR process. Cepstral features are much more resilient that spectral features used in missing data techniques. With the use of the combination of recognizers approach to ASR, supplementary information contained within cepstral domain extracted features can be stochastically coupled with the missing data spectral based features to achieve higher recognition performance. Through a series of experiments it has been established that the coupled approach will at least maintain existing recognition accuracy and will in fact increase recognition performance under all tested conditions. The coupled HMM model technique used in the combination of recognizers method, the fused HMM model, may be replaced with a more sophisticated coupling technique, the Mixed-Memory model which may yield an enhanced representation of the statistical relationship between HMMs. It is believed that such a model may further increase recognition performance and the incorporation of this coupled technique will be further investigated.

A known problem with the current experiment setup involves the means by which a noise estimate is obtained for ASR with missing data theory. Further investigations will entail enhancing this with the use of a silence detector and creating an adaptive noise estimate based upon a number of frames in each utterance processed.

5. REFERENCES

[1] A.C. Cooke, P.D. Green, L. Josifovski, and A. Vizinho, "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data," *Speech Communications*, pp. 267-285, 2001.

[2] B. Raj, R. Stern, "Missing-Feature Approaches in Speech Recognition," *IEEE Signal Processing Magazine*, pp. 101-116, Sept. 2005.

[3] 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.

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

[5] 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.

[6] J. Hakkinen, J, H. Haverinen, "On the Use of Missing Feature Theory with Cepstral Features," *Proceedings from Workshop on CRAC*, 2001.

[7] M.P. Cooke, A. Morris, P.D. Green, "Missing Data Techniques for Robust Speech Recognition," *ICSLP `94*, pp. 863 - 866, 1994.

[8] M. Brand, "Coupled Hidden Markov Models for Modeling Interacting Processes," Media Lab, MIT, 1997.

[9] L.K. Saul, M.I. Jordan, "Mixed memory Markov models: Decomposing complex stochastic process as mixtures of simpler ones," *Machine Learning*, vol. 37, pp. 75-87, 1999.

[10] 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*, Nov. 2005.

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

[12] J. Barker, "RESPITE CASA Toolkit CTK v1.1.1 User's Guide," University of Sheffield, 2001.