ROBUST SPEECH RECOGNITION USING MULTIPLE PRIOR MODELS FOR SPEECH RECONSTRUCTION

Arun Narayanan[*]*, Xiaojia Zhao*[*]*, DeLiang Wang and Eric Fosler-Lussier*

Department of Computer Science and Engineering
& Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
{narayaar, zhaox, dwang, fosler}@cse.ohio-state.edu

ABSTRACT

Prior models of speech have been used in robust automatic speech recognition to enhance noisy speech. Typically, a single prior model is trained by pooling the entire training data. In this paper we propose to train multiple prior models of speech instead of a single prior model. The prior models can be trained based on distinct characteristics of speech. In this study, they are trained based on voicing characteristics. The trained prior models are then used to reconstruct noisy speech. Significant improvements are obtained on the Aurora-4 robust speech recognition task when multiple priors are used; in conjunction with an uncertainty transform technique, multiple priors yield a 13.7% absolute improvement in the average word error rate over directly recognizing noisy speech.

Index Terms **—** Robust ASR, Aurora-4, uncertainty transform, feature reconstruction, CASA

1. INTRODUCTION

Robust recognition is one of the most challenging tasks facing automatic speech recognition (ASR) today. Traditional ASR methods perform well under clean speech conditions but lose performance in mismatched noise conditions. Some compensation strategies try to extract noise-robust features like RASTA and cepstral mean normalization (CMN), while others preprocess noisy speech using speech enhancement techniques. If noise samples are available *a priori*, models of speech and noise can be trained individually and used together during recognition. However, these approaches have subpar performance in real environments [3].

The fact that humans perform robust recognition with relative ease is attributed to auditory scene analysis (ASA) by Bregman [1]. Computational auditory scene analysis (CASA) tries to use perceptual cues to separate different sound sources. Several CASA based strategies have been proposed for robust ASR [2]. Most of these techniques make use of an estimated Ideal Binary Mask (IBM). An IBM is a binary mask with 1s and 0s representing speech dominated and noise dominated regions, respectively, of a noisy signal in the time-frequency (T-F) domain [2]. The true IBM can be created only if we have access to the clean speech signal and the noise that constitutes the noisy speech signal. In real conditions, the IBM has to be estimated directly from noisy speech. Estimated IBMs can be used in several ways: the probabilities of the unreliable parts (0s in the IBM) are marginalized in [3] to improve robust ASR performance. Alternatively, the unreliable parts can be reconstructed using prior models of speech [4]. This latter technique has the advantage of obtaining a complete speech feature which can then be transformed to the cepstral domain where ASR yields better results than the spectral domain. In an uncertainty transform technique, feature reconstruction and uncertainty decoding were used to obtain a significant improvement over a system that only uses reconstruction [5]. Uncertainty decoding accounts for uncertainty in reconstructed speech by adjusting the variance of the acoustic models.

In this paper, we explore the use of prior models of speech for reconstructing the unreliable components of a noisy utterance. Traditionally, such prior models are trained by pooling the entire speech data from the training set [2, 4, 5]. The fidelity of the reconstructed speech largely depends on how well the prior model is able to accurately match the unreliable components. Using a single speech prior model can be rather coarse, as speech characteristics can vary based on voicing characteristics, manner and place of articulation, etc. We investigate multiple speech priors, each of which models speech with a distinct characteristic. During the reconstruction stage, a segment of speech can be reconstructed based on its characteristic instead of using a single one-size-fits-all model. Such speech characteristics should also be detectable in noisy conditions with considerable accuracy for it to be useful during reconstruction. We expect that such a strategy can better reconstruct the unreliable components of a noisy utterance.

The rest of the paper is organized as follows: the next section provides the system description, followed by experimental results in section 3. We conclude with a discussion of results in section 4.

2. SYSTEM DESCRIPTION

In this section we describe our speech recognition system that uses multiple prior models. The prior models of speech are built based on the voicing characteristic (voiced vs. unvoiced) as it is an important, easily discernable characteristic of speech. Moreover, voiced/unvoiced (V/UV) detection can be performed with considerable accuracy even from noisy speech. The voiced and unvoiced speech prior models are trained independently. We begin with a description of the ASR process, followed by V/UV detection.

2.1. Recognition using multiple speech priors

A noisy utterance is first analyzed using FFT over 25 msec

[*] The first two authors contributed equally to this work.

windows with a window shift of 10 msec to produce 257 DFT coefficients for each frame. This creates a spectrogram for the noisy utterance. The next step is to identify reliable and unreliable components of the spectrogram. This was done using an IBM estimated directly from the noisy utterance. Although many sophisticated methods exist for IBM estimation [2], for the purpose of this study, we estimate IBM using a simple spectral subtraction method (see [5]).

Given an IBM, reliable and unreliable T-F units directly correspond to 1s and 0s, respectively, in the binary mask. Before we reconstruct the unreliable T-F units of a frame using the information available from the reliable ones, we need to identify whether the frame is voiced or unvoiced. This was done using a V/UV detector described in the next section. Once we know the voicing of the frame, we use the appropriate prior model to reconstruct the unreliable features of the frame.

Both voiced and unvoiced speech priors are modeled as a mixture of Gaussians as in [4, 5] - $p(X)$, where X corresponds to a random variable representing clean speech, is modeled as:

$$p(X) = \sum_{k=1}^{M} P(k)\, p(X \mid k) \tag{1}$$

Here, M corresponds to the number of Gaussians in the prior model, k the index, $P(k)$ the component weight and $p(X|k)$ the conditional probability density of X given the k^{th} Gaussian component. We use diagonal Gaussians in our experiments due to computational considerations. The difference from the model described in [5] is that we use two prior models, one for the voiced frames and one for the unvoiced frames, as mentioned before.

Using the reliable components, X_r, of speech, the unreliable components, X_u, are reconstructed by first estimating the *a posteriori* probability of the k^{th} Gaussian component, as shown in Equation (2), and then approximating the unreliable components as the expectation of X_u conditioned on X_r (Equation (3)) [5]:

$$P(k \mid X_r) = \frac{P(k)\, p(X_r \mid k)}{\sum_{l=1}^{M} P(l)\, p(X_r \mid l)} \tag{2}$$

$$\hat{X}_u = \sum_{k=1}^{M} P(k \mid X_r)\, \mu_{u,k} \tag{3}$$

$\mu_{u,k}$ here refers to the unreliable components of the mean vector of the k^{th} Gaussian in the speech prior. The signal is then re-synthesized from the reconstructed features and converted to the cepstral domain to obtain enhanced MFCC features for the utterance.

The spectral uncertainties in the estimate of the reconstructed features are approximated as in [5]:

$$\hat{\sigma}^2 = \sum_{k=1}^{M} P(k \mid X_r) \left\{ \left(\begin{bmatrix} X_r \\ \hat{X}_u \end{bmatrix} - \mu_k \right)^2 + \begin{bmatrix} 0 \\ \sigma_{u,k}^2 \end{bmatrix} \right\} \tag{4}$$

Here σ and $\hat{\sigma}$ refer to the standard deviation of the Gaussians in the prior model and the estimated standard deviation, respectively. The estimated spectral variance is then transformed to the cepstral domain using a multilayer perceptron (MLP) as described in [5]. In order to obtain cepstral variance, apart from the spectral variance, the MLP takes as input the enhanced cepstral values corresponding to that frame, a preceding frame and a succeeding frame [5].

The above equations are a general treatment to using prior models of speech for feature reconstruction; but the proposed method handles voiced speech and unvoiced speech separately using two independently trained prior models. Using V to denote the voicing of a frame (1 being voiced and 0 unvoiced) Equations (2) and (3) can be modified to accommodate multipriors as:

$$P(k_V \mid X_r) = \frac{P(k_V)\, p(X_r \mid k_V)}{\sum_{l_V=1}^{M_V} P(l_V)\, p(X_r \mid l_V)}\, f_V(X); V \in \{0,1\} \tag{5}$$

$$\hat{X}_u = \sum_{V \in \{0,1\}} f_V(X) \sum_{k_V=1}^{M_V} P(k_V \mid X_r)\, \mu_{u,k_V} \tag{6}$$

where $f_V(X)$ is an indicator function that is 1 iff V matches the output of the V/UV detector (V=1 implies voiced, 0 unvoiced). The spectral variance can also be estimated in a similar fashion.

Once we obtain the enhanced cepstral features and the estimated uncertainties, recognition is performed using a traditional HMM-based ASR system trained on clean speech. The estimated variances were used to adjust variances of the trained acoustic models during the decoding stage [5].

2.2. Voiced/unvoiced detection

Voiced/unvoiced detection is modeled as a binary decision problem. Gaussian mixture models (GMM) are trained independently to model noisy voiced speech and noisy unvoiced speech. The V/UV decisions are made at the frame level by comparing the log likelihood ratio (LLR) of the data relative to the voiced and unvoiced models; the LLR is compared to a threshold, τ, as shown in Equation (7):

$$\log\left(\frac{p(X \mid V=1)}{p(X \mid V=0)} \right) \geq \tau \tag{7}$$

X here refers to the noisy observation and V denotes the assumed voicing of the noisy frame (1=voiced, 0=unvoiced). The threshold is set to 1, instead of 0, to reduce the false alarm errors and thereby improve the V/UV detection accuracy. The likelihoods are estimated by an equation very similar to Equation (1), but with an additional conditioning on the model (voiced or unvoiced) under which the likelihood needs to be estimated. We also post-process the output to re-label spurious isolated voiced or unvoiced frames: the detected voicing of such an isolated frame is switched if both its neighbors have voicing characteristics different from its own.

An alternative way to perform V/UV detection is to binarize the outputs of pitch detectors like Praat [6] or other noise robust pitch detection algorithms (PDA) [7]. However, such PDAs focus on estimating the correct pitch values in each frame. Although they work well in clean conditions, they are likely to have difficulties dealing with noisy speech especially if the underlying noise has a harmonic structure. We overcome this by training our GMMs to focus solely on V/UV detection directly from noisy speech, as we believe that this will yield better V/UV detection accuracies. A systematic comparison between our approach and a Praat based V/UV detection system is presented in the next section.

3. RESULTS

3.1. Experimental setup

The proposed ASR system is evaluated on the Aurora-4, 5000 word closed-vocabulary recognition task [8]. The Aurora-4 corpus

is based on the *Wall Street Journal* (WSJ0) database [9]. It consists of clean speech utterances digitally mixed with different noise types at SNR levels ranging from 5 to 15 dB. Although the corpus also has recordings that simulate different conditions like variations in microphone and sampling frequency, we only use the noisy samples from the corpus, sampled at 16 kHz, as our primary goal is to improve noise robustness [8].

Clean speech utterances from the training set are used to train the HMM based speech recognizer using the HTK Toolkit [10]. The training set consists of 7138 utterances. Cepstral mean normalized Mel frequency cepstral coefficients (MFCC) along with their delta and acceleration coefficients and normalized log energy (MFCC_E_D_A_Z in HTK terminology) are used. The models themselves consist of state-tied cross-word triphone-based HMMs. The observation density of each state is modeled as a mixture of 16 diagonal Gaussian components. The standard bigram language model and the CMU pronunciation dictionary-based lexicon are used. The reduced test set, which consists of 166 utterances for each noise type, is used to evaluate the proposed method.

The prior models of speech, implemented as large GMMs, are also trained using the HTK toolkit. The GMMs are trained using spectral features. Both voiced and unvoiced models consist of 1024 diagonal Gaussian components.

The reliable and unreliable components of speech are identified using an estimated IBM. To estimate the IBM, the spectra of the first and last 50 frames of a noisy utterance are averaged to create an estimate of the spectrum of noise. Using the noise spectrum, the local SNR at each T-F unit is estimated. A T-F unit is then labeled 1 if the estimated local SNR is above a threshold and 0 otherwise. The threshold was set to 5 dB as suggested in [5]. The estimated noise energy is subtracted from the mixture energy before the unreliable components are reconstructed [4, 5].

We randomly pick 40 noisy utterances from the multi-noise training set of Aurora-4 to train the MLP that transforms spectral uncertainties to cepstral uncertainties. The IBM, created using the clean speech signal corresponding to those noisy utterances and the noise signal estimated by subtracting the clean speech signal from the mixture signal, is used to estimate the spectral uncertainty. The target is the true variance of the enhanced cepstral coefficients with respect to the clean cepstral coefficients. The MLP has 374 input units (257 spectral uncertainties + 39x3 cepstral coefficients), 800 hidden units and 39 output units [5]. The transfer function used for the hidden units is the tangential sigmoid function and that of the output units is the linear function. The MLP was trained for 150 iterations to minimize the mean squared error.

Noisy utterances (2676 in number) from the multi-noise training set of Aurora-4 are used to train the V/UV detector. Pitch is estimated from clean speech utterances corresponding to these noisy utterances using Praat and binarized to establish the ground truth V/UV labeling for the training set. MFCC coefficients (MFCC_E_D_A_Z) extracted from noisy utterances are used to train the GMMs corresponding to the voiced model and the unvoiced model. Each model has 1024 diagonal Gaussian components. Also, since the first and last 50 frames are used to estimate the noise spectrum during the IBM estimation, they are labeled as unvoiced irrespective of the output obtained using the V/UV detector.

3.2. Experimental results

The V/UV detection accuracies are shown in Table 1 for the noise types in Aurora-4. The table shows the percent accuracies of the detector. As a comparison, V/UV intervals were also estimated from noisy speech directly using Praat. Pitch estimated from clean speech using Praat is used to establish the ground truth V/UV label for each frame. Clearly, the GMM based modeling does a much better job in identifying voiced and unvoiced frames. Note that the first and last 50 frames are labeled as unvoiced even for the Praat based V/UV detector.

Table 1. V/UV detection accuracies of the proposed GMM based method and the Praat based method. The last row shows the average accuracy.

Noise Type	GMM	Praat
Car	87.0%	52.0%
Babble	85.8%	72.0%
Restaurant	86.8%	74.0%
Street	86.4%	62.5%
Airport	86.5%	71.5%
Train	85.4%	63.2%
Average	86.3%	65.9%

The ASR word error rates (WER) calculated as the sum of substitution, insertion and deletion errors divided by the total number of words are shown in Table 2. When tested on clean speech, a WER of 8.7% was obtained. The first row in the table shows the baseline results. They are obtained when features are directly extracted from the noisy utterances without any enhancement, apart from CMN. The first set of results is obtained when the unreliable components are reconstructed but ASR is performed without using the uncertainty transform technique. As a baseline for the reconstruction based methods, features were enhanced using only a single prior (SP) model, as in [5]. The results are shown in the second row of the table. Note that the SP model was trained in a setting similar to the proposed model to make the results comparable. The next three rows show results using multiprior (MP) models. The V/UV detector is implemented by either using Praat (MP-Praat) or Gaussian mixture models (MP-GMM) or by using the ground truth V/UV (MP-GT) information based on the binarized pitch values. As Praat does not identify the V/UV frames that well, the WERs for MP-Praat are better only in some cases when compared to SP. But for MP-GMM and MP-GT, the WER is better than SP for all noise types. The improvements are statistically significant for car noise, babble noise, street noise and train noise.

The next set of results show WERs when the uncertainty transform technique is included during the decoding stage (SP-UT, MP-Praat-UT, MP-GMM-UT, MP-GT-UT, following the same order as before). As can be seen, in almost all the cases, there is an improvement when the uncertainty transform is included. MP-GT-UT for restaurant noise and train noise are the exceptions, but the drop in performance is not statistically significant. The improvements for MP-GMM-UT are statistically significant for street noise and train noise when compared to SP-UT. The least improvements in performance were obtained for restaurant noise and airport noise. This is partly because of the impulsive nature of these noise types, which severely affects the quality of the spectral subtraction mask. Note that even for the single prior based method the least improvements were obtained for these two noise types, when compared to the baseline method.

Table 2. WER of different recognition methods, as discussed in the 'Experimental results' section for the 6 noise types in Aurora-4. SP and MP abbreviate single prior and multiprior, respectively, indicating the number of prior models used for feature reconstruction. Praat, GMM (Mixture of Gaussians) and GT (Ground Truth) refer to the strategy used for V/UV detection. The suffix UT denotes the use of the uncertainty transform technique during recognition. The last column shows the average WER of each of the systems across all noise types.

System	Test Set						
	Car	Babble	Restaurant	Street	Airport	Train	Average
Baseline	44.9	43.7	43.2	52.0	44.1	55.2	47.2
SP	21.5	38.5	42.6	41.5	41.5	39.4	37.5
MP-Praat	21.6	36.6	42.8	41.5	42.5	38.8	37.3
MP-GMM	**19.6**	**34.8**	**41.0**	**38.3**	**41.1**	**36.5**	**35.2**
MP-GT	17.9	34.4	41.4	37.7	39.8	35.0	34.4
SP-UT	18.9	34.2	41.2	40.6	37.0	39.0	35.2
MP-Praat-UT	19.4	33.8	39.7	40.5	37.1	38.8	34.9
MP-GMM-UT	**18.4**	**32.8**	**39.1**	**37.4**	**36.9**	**36.5**	**33.5**
MP-GT-UT	16.6	32.7	40.1	37.4	36.6	35.5	33.2

The last column shows the average performance of these methods across all noise types. By using multiple priors, we can obtain a better average performance even when V/UV detection is made using Praat. The average improvement of the multiprior based method, when the proposed V/UV detector or the ground truth based detector is used, is statistically significant as compared to the single prior based method. This is true irrespective of whether the uncertainty transform is used during the decoding stage of the recognizer. In most cases, there is still some improvement to be gained when an ideal detector is used for V/UV classification. This would mean that improvements in V/UV detection can further improve the performance of the proposed method.

4. DISCUSSIONS

We have shown that using multiple speech priors based on the voicing characteristic of speech can be advantageous for robust automatic speech recognition when speech priors are used to reconstruct features corrupted by noise. The improvement holds even when techniques like uncertainty transform are used to account for the distortions in the reconstructed speech. We believe that multipriors give a better representation of speech as they are built to represent different categories of speech. One of the possible disadvantages of using multipriors is that we have to first predict the characteristic of each frame of speech before we can reconstruct corrupted components. In this study, this is overcome by building a simple yet effective voiced/unvoiced detector using Gaussian mixture models.

We believe that the proposed strategy opens up new avenues to improve robust speech recognition. Voicing of speech is only one of many possible characteristics that can be utilized under the proposed framework. Other examples include place and manner of articulation, and gender based prior models. These will be explored in future research.

Finally, we would like to point out that identifying the reliable and unreliable components of speech is an important subtask that has major implications to the quality of reconstructed speech and in turn, recognition accuracies. Although we use a simple spectral subtraction mask for our study, we can expect better performance if more sophisticated CASA based strategies are used [2].

Acknowledgements. We would like to thank Jeremy Morris, William Hartmann, and Srinivasan Soundararajan for providing the modified HTK training and decoding scripts. The research described in this paper was supported in part by an AFOSR grant (FA9550-08-1-0155).

5. REFERENCES

[1] A.S. Bregman, *Auditory Scene Analysis,* MIT Press, Cambridge, MA, 1990.

[2] D.L. Wang, G.J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications,* Wiley/IEEE Press, Hoboken, NJ, 2006.

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun*., vol. 34, pp. 267–285, 2001.

[4] B. Raj, M.L. Seltzer, and R.M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, pp. 275–296, 2004.

[5] S. Srinivasan, and D.L. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE T-ASLP,* vol. 15, pp. 2130-2140, 2007.

[6] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (Version 5.0.02)", Available: http://www.fon.hum.uva.nl/praat, Dec 2007.

[7] Z. Jin and D.L. Wang, "A multipitch tracking algorithm for noisy and reverberent speech," *Proc. ICASSP*, pp. 4218-4221, 2010.

[8] N. Parihar and J. Picone, "Analysis of the Aurora large vocabulary evalutions," *Proc. Eurospeech*, pp. 337–340, 2003.

[9] D. Paul and J. Baker, "The design of Wall street journal-based CSR corpus," Proc. ICSLP, pp. 899–902, 1992.

[10] S. Young, D. Kershaw, J. Odell, V. Valtchev, and P. Woodland*, The HTK Book (for HTK Version 3.4)*, Microsoft Corp., Redmond, WA, 2009.