CLASSIFIER-BASED MASK ESTIMATION FOR MISSING FEATURE METHODS OF ROBUST SPEECH RECOGNITION

Michael L. Seltzer, Bhiksha Raj, and Richard M. Stern

Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213, USA
{mseltzer, bhiksha, rms}@cs.cmu.edu

ABSTRACT

Missing feature methods of noise compensation for speech recognition operate by removing components of a spectrographic representation of speech that are considered to be corrupt, as indicated by a low signal-to-noise ratio. Recognition is either performed directly on the incomplete spectrograms or the missing components are reconstructed prior to recognition. These methods require a spectrographic mask which accurately labels the reliable and corrupt regions of the spectrogram. Current methods of mask estimation rely on assumptions about the corrupting noise such as stationarity. This is a significant drawback since the missing feature methods themselves have no such restrictions. We present a new mask estimation technique that uses a Bayesian classifier to determine the reliability of spectrographic elements. Features were designed that make no assumptions about the corrupting noise signal, but rather exploit characteristics of the speech signal itself. Missing feature compensation experiments were performed on speech corrupted by a variety of noises. In all cases, classifier-based mask estimation resulted in significantly better recognition accuracy than conventional mask estimation methods.

1. INTRODUCTION

When speech is corrupted by noise, speech recognition accuracy degrades, especially when the recognition system has been trained on clean speech (*e.g.* [4]). There have been many algorithms proposed that compensate for the negative effects of noise in speech and greatly improve recognition accuracy. However, these methods assume that the corrupting noise is stationary. If the noise is non-stationary, these methods fail.

The missing feature methods are a promising new group of robustness techniques for compensating for non-stationary noise. The missing feature paradigm is based on the notion that noise affects different time-frequency regions of speech differently. In a spectrographic display of speech, there will be regions of low SNR and high SNR depending on the relative energies of the speech and the noise at each time-frequency location. In missing feature methods, regions with low SNR are considered "corrupt" or "missing" and are removed from the spectrogram. Noise compensation is performed either by reconstructing the missing elements from the remaining reliable regions prior to recognition, or by performing recognition directly on the incomplete spectrograms. Unlike other compensation methods, these techniques require no assumptions about the corrupting noise signal such as stationarity. They do, however, require a spectrographic mask which accurately labels every time-frequency location as reliable or corrupt. Missing feature methods have been shown to be very successful at compensating for stationary and non-stationary noise when this spectrographic mask is completely known *a priori*. However, when the masks are unknown, these techniques are unusable.

Clearly then, estimating spectrographic masks is of critical importance to the success of missing feature methods. Current methods of spectrographic mask estimation rely on a running estimate of the noise spectrum obtained via spectral subtraction to estimate the local SNR at each time-frequency location [2]. The SNR estimates are compared to a specified threshold and those below the threshold are considered corrupt while those above it are considered reliable.

Such mask estimation methods perform well when the corrupting noise is stationary, as this assumption is required for spectral subtraction. However, when the noise is non-stationary, masks estimated in this manner are very inaccurate. This is illustrated in Figure 1. Missing feature compensation has been applied to noisy speech using masks estimated using spectral subtraction and "oracle" masks generated from full *a priori* knowledge of the noise signal. Figure 1a shows recognition accuracy vs. SNR for speech that has been corrupted with white noise. There is significant improvement over baseline accuracy using spectral subtraction to estimate the masks. Figure 1b shows the same plot for speech corrupted by music, which is highly non-stationary. Here, spectral subtraction completely fails in mask estimation. In fact, recognition accuracy after compensation using these masks is slightly worse than the baseline uncompensated recognition. However, the accuracy obtained using oracle masks in both plots

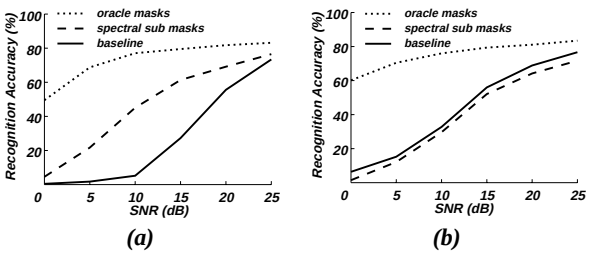

Figure 1. Recognition accuracy vs. SNR when missing feature compensation is applied to noise corrupted speech. (a) Speech corrupted by white noise. (b) Speech corrupted by music.

show the potential of missing feature methods for noise compensation if the masks can be estimated reliably.

In this paper we present a new mask-estimation technique that uses a Bayesian classification strategy to determine the reliability of each spectrographic element. Casting mask estimation as a classification problem has three distinct advantages. First, the problem of mask estimation is reduced from the difficult task of SNR estimation to a simpler binary decision process. Second, the classification scheme allows any information that is pertinent to be incorporated into the mask estimation decision. Finally, with an appropriate choice of features, mask estimation can be free of assumptions about the corrupting noise.

In Section 2 we describe the feature set used by the classifier to estimate the spectrographic masks. In Section 3 we describe the classification strategy we use. We describe experiments that were performed to test the mask estimation strategy in Section 4, and in Section 5 we summarize our findings.

2. FEATURE EXTRACTION

Because voiced speech and unvoiced speech are generated by different production mechanisms, they have very different characteristics. As a result, we make a distinction between features used to classify the reliability of spectrographic locations of voiced speech and those used for regions of unvoiced speech. Additionally, because mask estimation should be free of assumptions about the corrupting noise, we designed a feature set that exploits the inherent characteristics of the speech signal itself.

2.1. Features for voiced speech

Two key characteristics of voiced speech that we exploit are the presence of a strong fundamental frequency and its harmonics, and a distinctive spectral contour across frequency.

2.1.1. Comb Ratio

Because of the harmonic nature of voiced speech, the majority of the energy of a clean voiced speech signal resides in its harmonics [5]. Additive noise does not typically have this characteristic. When additive noise is mixed with voiced speech, the overall signal energy increases both at the harmonics of the pitch and at the frequencies in between. Therefore, a measure that compares the energy at the harmonics of voiced speech to the energy outside the harmonics is a good indicator of noise present in the signal.

Of course, such a measure requires a pitch estimator that is robust to noise. Our pitch estimation algorithm is based on a multi-band analysis of speech [8]. Each frame of speech is passed through a bank of bandpass filters. At each filter output, a pitch estimate is computed from the autocorrelation of the frame. For voiced speech, a single pitch dominates the distribution of the candidate pitch values. However, for unvoiced speech, the distribution is roughly uniform. The overall pitch estimate, F_0, for the frame is determined by majority rule: if a single pitch dominates 25% or more of the frequency bands, the frame is assumed to be voiced; otherwise, it is considered unvoiced. This method is similar to that in [3], except that the pitch estimate in [3] is obtained by pooling the autocorrelations obtained in the various bands together, rather than on the basis of a majority rule of the individual pitch estimates.

The pitch estimate is used to construct a comb filter that captures the energy present in the harmonics of voiced speech. We use an IIR comb filter implementation given by the transfer function in Equation (1), where $p = 1/F_0$ is the pitch period and g is a tunable parameter which sets the sharpness of the teeth of the comb.

$$H_{comb}(z) = z^{-p}/(1 - gz^{-p}) \quad (1)$$

It was determined empirically that setting $g = 0.7$ captures most of the harmonic information of voiced speech.

To capture the energy of the components of the signal that fall in between the harmonics, the comb filter is simply shifted by $F_0/2$. The transfer function for this shifted comb filter is given by Equation (2).

$$H_{combshift}(z) = -z^{-p}/(1 + gz^{-p}) \quad (2)$$

If we assume that the voiced speech resides at the harmonics of the fundamental frequency while noise may reside in all frequency bands, the energy at the output of the comb filter is a measure of speech and noise energy while that of the shifted comb filter is a measure of noise energy only. Thus, the log ratio of the energies of the speech signal passed through the comb and shifted comb filters is a measure of speech plus noise to noise. The cleaner the speech signal is, the larger this ratio will be. We call this feature the *comb ratio*. The comb ratio, $CR(n, \omega_i)$, is given by Equation (3), where y_{comb} and $y_{combshift}$ are the outputs obtained after the speech signal in frame n and subband ω_i has been passed through the comb and shifted comb filters, respectively.

$$CR(n, \omega_i) = 10\log_{10}\left(\frac{\sum_k y_{comb}[k, \omega_i]^2}{\sum_k y_{combshift}[k, \omega_i]^2}\right) \quad (3)$$

2.1.2. Autocorrelation Peak Ratio

Voiced speech is a quasi-periodic signal. The secondary peaks in the autocorrelation function of a frame of voiced speech will be less than or equal to the height of the main peak. The less periodic the signal is, the smaller the secondary peaks will be. Adding uncorrelated noise to a signal effectively reduces its periodicity, increasing the difference in the heights of the main peak and the secondary peaks. We use the ratio of the height of largest secondary peak to the height of the main peak as a measure of periodicity. This autocorrelation peak ratio feature will be close to one for clean speech and decrease as the signal is increasingly corrupted by noise.

2.1.3. Subband Energy to Fullband Energy Ratio

In addition to its characteristic harmonicity, voiced speech has a distinct spectral shape. The energy of voiced frames is concentrated at the lower frequencies and tails off at higher frequencies. As noise is added to the speech, its spectral shape will change as a function of the spectral characteristics of the noise. The log ratio of the energy in a subband to the overall frame energy captures the effect of additive noise on a particular subband and on the overall contour.

2.1.4. Subband Energy to Subband Noise Floor Ratio

Having knowledge of the noise floor of a noise-corrupted speech signal is obviously very useful for estimating the SNR. However, an accurate measure of the noise floor is difficult to obtain. If we assume that the corrupting noise is stationary, we can coarsely estimate the level of the noise floor in a particular subband by looking at the distribution of the energy in that subband across all frames in an utterance. These distributions typically have two modes, one at a low energy value representing the silence and low energy speech regions and one at a higher energy representing high energy speech regions. The idea of statistically modeling the energy distributions of speech has been used for speech endpoint detection using HMMs [1]. We have used a much simpler technique to get a rough estimate of the noise floor. The energies of all frames of an utterance are put into a histogram and the lower energy peak is found. The energy bin in the histo-

gram corresponding to this peak value is considered the noise floor of the noisy speech signal. We use the ratio of the energy in a subband of a frame of speech to the estimate of the noise floor in that subband of the utterance as a feature to help determine the likelihood that a specific spectrographic location has been corrupted by noise. We note that this technique is similar to spectral subtraction in that we are using the energy of the silence frames to estimate the noise floor of the entire utterance. If the noise is highly non-stationary, the noise floor estimate will not necessarily be accurate.

2.1.5. Flatness

As was noted earlier, voiced speech exhibits a very definitive trajectory across frequency, and when noise is added to speech, this spectral shape will change. The valleys in the spectrum tend to flatten as noise is added to a speech signal. This "flatness" can be characterized by the variance of the subband energy in a neighborhood of spectrographic locations around a given pixel. For a given subband, a signal corrupted with noise tends to have shallower, flatter valleys than its uncorrupted counterpart. Therefore, we expect noise-corrupted spectrographic locations to have a lower variance than cleaner ones.

2.2. Features for Unvoiced Speech

Unvoiced speech is much more difficult to characterize than voiced speech. There is no harmonicity or other regularity as in voiced speech. As a result, the pitch-related features developed for voiced speech will be ineffective for unvoiced speech. Unvoiced speech also has lower energy than voiced speech and is therefore more affected by noise than voiced frames. However, it does have a general spectral shape that is unlike voiced speech and most naturally occurring noises. Unvoiced speech energy is concentrated at the higher frequencies and tails off at lower frequencies. The three voiced speech features that do not rely on pitch characterize a frame of speech in terms of the relative energy levels in each of the subbands, and the overall and local spectral shape. They are useful features because we know that adding noise to a speech signal alters both the relative subband energy levels and the spectral shape. This is true for both voiced and unvoiced speech. While the energy distribution of unvoiced speech across frequency is very different from that of voiced speech, it too will be altered by additive noise. As a result, we can use the remaining three non-pitch dependent features (subband-to-fullband energy ratio, subband energy to subband noise floor ratio, and flatness) to characterize unvoiced speech.

3. CLASSIFICATION STRATEGY

A multivariate Gaussian classifier with a full covariance matrix was used for mask estimation. Each pixel was represented by a feature vector of length five or three, depending on whether the frame was voiced or unvoiced. Because of the differing feature sets, separate classifiers were constructed for voiced frames and unvoiced frames. In addition, the feature values themselves may differ significantly from subband to subband *within* each class. Therefore, we also implemented a separate classifier for each subband.

Missing feature algorithms treat spectrographic elements below a certain SNR as missing or corrupt and effectively remove them from the spectrogram. This SNR threshold, which varies depending on the missing feature method applied, is used to label the data used to train the mask estimation classifier.

The prior probabilities of a reliable and corrupt element were determined using a cross-validation data set. We expect the prior probabilities of corrupt and reliable spectrographic elements to vary with the global SNR, as more elements are corrupt at higher noise levels than at lower noise levels. However, because we did not know the global SNR, we chose the constant prior probabilities that yield the best recognition accuracy over all SNRs.

4. EXPERIMENTAL RESULTS

Experiments in classifier-based mask estimation were performed using the DARPA Resource Management (RM1) corpus [6], corrupted by three different noise environments: stationary white noise, factory noise, consisting of quasi-stationary background noise mixed with non-stationary impulsive noises, and music from the "Marketplace" radio program, which is highly non-stationary.

To form a complete missing feature compensation system, classifier-based mask estimation was combined with the cluster-based missing feature reconstruction algorithm [7]. In cluster-based reconstruction, log Mel spectral vectors from clean speech are clustered. Missing features from noise-corrupted speech, identified by the spectrographic mask, are recovered by first identifying the closest cluster based on the values of the features that are present, and then estimating the missing values using MAP procedures. The reconstructed log Mel spectral vectors are then transformed to standard Mel frequency cepstra for recognition. This algorithm performs optimally with a corrupt/reliable SNR threshold of –5 dB.

For each noise environment (white noise, factory, music), the following experimental procedure was followed:

The classifier was trained on 2880 utterances from RM1, corrupted with noise to various SNRs. For training, the pitch estimates required for the pitch-dependent classifier features were estimated from clean speech using the method described in Section 2. The local SNR was computed for every spectrographic element, and the training data were labelled accordingly. The means and covariance matrices of the classifier were estimated for each subband and for each type of speech. A cross-validation data set of 200 utterances from RM1 was used to determine the prior probabilities. There was no overlap between the cross validation, training, and test sets. Based on the cross validation data, the prior probability of the reliable elements was set to 0.8 for the white noise and factory noise environments and 0.6 for the music environment.

The test set consisted of 1600 utterances from RM1. The pitch estimates for the test set were derived directly from the noisy speech for the white noise and factory noise experiments. Because music is highly harmonic, the pitch detection algorithm performed poorly on speech corrupted by music, so pitch estimates from clean speech were used for this case. The spectrographic masks of the noise-corrupted speech were estimated by the classifier. No information other than overall environment (*i.e.* white noise, factory noise, or music) such as local or global SNR, was known to the classifier.

The accuracy of the classifier was measured by comparing the estimated masks to oracle masks known to be correct. These oracle masks, generated from full *a priori* knowledge of the noise signal, represent the best possible spectrographic masks. Tables 1 and 2 show confusion matrices for the classifier for speech in

the three noise environments for voiced speech and unvoiced speech, respectively. Reliable elements are labelled as Class 1 and corrupt elements are labelled as Class 0.

	AWGN		Factory		Music	
	"1"	*"0"*	*"1"*	*"0"*	*"1"*	*"0"*
1	87%	13%	79%	21%	72%	28%
0	16%	84%	21%	79%	33%	67%

Table 1: Classifier accuracy for voiced frames for speech in three noise environments. Reliable elements are Class 1 and corrupt elements are Class 0.

	AWGN		Factory		Music	
	"1"	*"0"*	*"1"*	*"0"*	*"1"*	*"0"*
1	76%	24%	71%	29%	64%	36%
0	13%	87%	22%	78%	28%	72%

Table 2: Classifier accuracy for unvoiced frames for speech in three noise environments. Reliable elements are Class 1 and corrupt elements are Class 0.

To perform missing feature compensation, cluster-based reconstruction was performed on the spectrographic elements identified as corrupt by the estimated masks. Recognition was performed using the SPHINX-III speech recognition system. Context-dependent continuous HMMs (1 Gaussian/state) were trained on clean speech using 2880 sentences from RM1. No delta or double delta cepstra were used.

The recognition results are shown in Figure 2. As the plots clearly indicate, classifier-based spectrographic masks resulted in significantly better recognition accuracy than spectral subtraction-based masks in all three noise environments.

5. SUMMARY

In this paper we have presented a new method of spectrographic mask estimation for missing feature compensation. We recast mask estimation from an SNR estimation problem to a Bayesian classification problem. In doing so we have been able to remove the stationarity limitations that the previous mask estimation methods placed on the corrupting noise signal by creating a unique feature set that exploits the inherent characteristics of the speech signal itself. We demonstrate that masks generated by the classifier result in significantly better recognition accuracy than masks produced using conventional noise estimation methods. While the classifier makes no assumptions about the corrupting noise signal, it does currently require knowledge of the overall operating environment. However, we do not feel that this is a serious limitation, as this information is readily available in most situations.

Missing-feature methods for noise compensation in speech recognition are gaining popularity both because they are capable of significant improvements in recognition accuracy and because they make logical sense based on our knowledge of the human auditory system. Similarly, building a classifier that uses features that are based on the intrinsic characteristics of the speech signal itself is also intuitively satisfying. Because no assumptions about

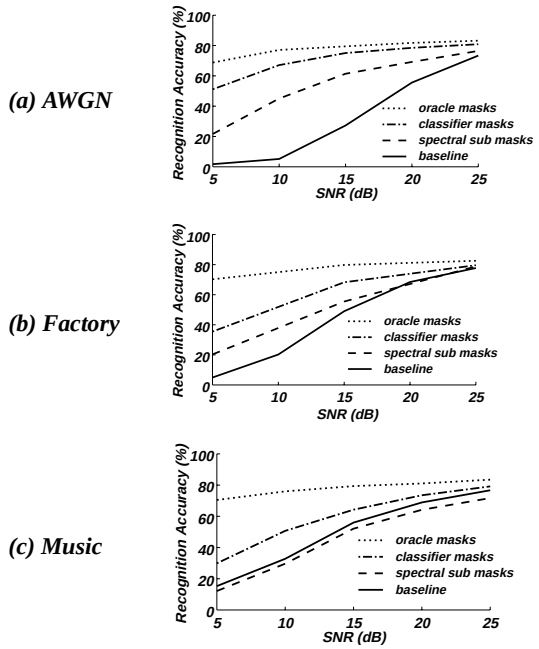

Figure 2. Recognition accuracy vs. SNR for speech in three different noise environments. Missing feature compensation was performed used masks estimated by the classifier, spectral subtraction, or full *a priori* knowledge of the noise (oracle).

the noise are made, it is logical that the classifier will be able to estimate spectrographic masks for many, if not all, noise types. This is a significant improvement over previous mask estimation methods.

ACKNOWLEDGEMENTS

This research was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

REFERENCES

[1] Acero, A., Crespo, C., and Torrecilla, J.C., "Robust HMM-based endpoint detector," *Proc. Eurospeech '93*, p. 1551-1554.

[2] Cooke, M., Green, P., Josifovski, L., and Vizinho, A., "Robust ASR with Unreliable Data and Minimal Assumptions," *Proc. of Workshop on Robust Methods for Speech Recognition in Adverse Conditions '99*.

[3] Meddis, R., and Hewitt, M.J. , "Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. Pitch identfication," *Journal of the Acoustical Society of America*, vol. 89 no. 6, June, 1991, p. 2866-2882.

[4] Moreno, P. J., Raj, B., and Stern, R. M., "Data driven environmental compensation for speech recognition: a unified approach" *Speech Communication*, no. 24, 1998, p.267-285.

[5] Morgan, D.P., George, E.B, Lee, L.T., and Kay, S.M., "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. on Speech and Audio Processing*, vol. 5 no. 5, Sept. 1997, p. 407-424.

[6] Price, P., Fisher, W.M., Bernstein, J., and Pallet, D.S., "The DARPA 1000 word Resource Management database for continuous speech recognition," *Proc. ICASSP '88*, p. 651-654.

[7] Raj, B., Seltzer, M.L., and Stern, R.M., "Reconstruction of damaged spectrographic regions for robust speech recognition," *Proc. ICSLP '00*.

[8] Seltzer, M.L., and Stern, R.M., "Histogram-Based Robust Pitch Estimation", submitted to *IEEE Signal Proc. Letters*.