Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 43 (2004) 379–393
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition

Michael L. Seltzer [a,*], Bhiksha Raj [b], Richard M. Stern [c]

[a] *Microsoft Research, Redmond, WA 98052, USA*
[b] *Mitsubishi Electric Research Labs, Cambridge, MA 02139, USA*
[c] *Department of Electrical and Computer Engineering and School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

Received 21 May 2002; received in revised form 9 May 2003; accepted 31 March 2004

Abstract

Missing feature methods of noise compensation for speech recognition operate by first identifying components of a spectrographic representation of speech that are considered to be corrupt. Recognition is then performed either using only the remaining reliable components, or the corrupt components are reconstructed prior to recognition. These methods require a spectrographic mask which accurately labels the reliable and corrupt regions of the spectrogram. Depending on the missing feature method applied, these masks must either contain binary values or probabilistic values. Current mask estimation techniques rely on explicit estimation of the characteristics of the corrupting noise. The estimation process usually assumes that the noise is pseudo-stationary or varies slowly with time. 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 used for classification were designed that make no assumptions about the corrupting noise signal, but rather exploit characteristics of the speech signal itself. Experiments were performed on speech corrupted by a variety of noises, using missing feature compensation methods which require binary masks and probabilistic masks. In all cases, the proposed Bayesian mask estimation method resulted in significantly better recognition accuracy than conventional mask estimation approaches.

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. Moreno, 1996). There have been many algorithms proposed that compensate for the

* Corresponding author. Address: Microsoft Research, 1 Microsoft Way, Redmond, WA 98052, USA. Tel.: +1 425 706 3763; fax: +1 425 936 7329.
E-mail addresses: mseltzer@microsoft.com (M.L. Seltzer), bhiksha@merl.com (B. Raj), rms@cs.cmu.edu (R.M. Stern).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.03.006

negative effects of noise in speech and greatly improve recognition accuracy. However, many of these methods assume that the corrupting noise is stationary or slowly varying. If this assumption is violated, these methods perform poorly.

Missing feature methods (e.g. Raj et al., 2000; Raj et al., 1998; Cooke et al., 2001; Renevey, 2000) are a group of techniques developed over the last several years for compensating for additive noise, regardless of its stationarity. The missing feature paradigm is based on the notion that noise affects different time-frequency regions of speech differently. In a spectrographic display of noisy 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. Regions with low SNR are considered "corrupt" while regions with high SNR are dubbed "reliable". In conventional missing feature approaches, the low SNR components are deemed unreliable and disregarded (hence, "missing"). Missing feature compensation techniques operate on this incomplete spectrogram either by estimating the proper values of unreliable components and then performing recognition on the now-complete sequence of feature vectors (Raj et al., 2000), or by passing the incomplete feature vectors directly to a recognition system which has been modified to operate on partial vectors (Cooke et al., 2001). We refer to the former family of methods as *feature-compensation methods* and the latter family as *classifier-compensation methods* (Raj et al., 2001).

Unlike other compensation methods, these techniques require no assumptions about the corrupting noise signal, e.g. stationarity. However, all missing feature approaches do require all components in a spectrographic display of speech to have a labelling describing their degree of "reliability" or "corruption". We refer to such a labelling as a *spectrographic mask*. Conventional missing feature methods require a binary tagging of spectrographic locations, typically using a local SNR criterion. That is, elements below a given SNR are tagged as corrupt (0), while those above are labelled reliable (1). This labelling for all elements in an utterance is captured in a binary spectrographic mask.

More recent missing feature methods, such as Barker et al. (2000), have shown improved performance using soft decisions in the mask estimation process. The use of binary masks forces a hard decision to be made about whether each element is dominated by speech or by noise. In contrast, the label assigned to each spectrographic element by a soft-decision mask can take on a continuum of values between 0 and 1. This label can be interpreted as the probability that a particular element is dominated by speech. Elements with a mask probability approaching 1 have strong evidence that they contain primarily speech and very little noise, while those elements with a mask probability approaching 0 are strongly believed to contain primarily noise and very little speech.

Missing feature methods have been shown to be very successful at compensating for the effects of stationary and non-stationary noise when this mask is computed from a priori knowledge of the SNR of all spectrographic components. However, when the masks are unknown, these techniques are unusable.

Clearly then, reliable estimation of spectrographic masks is of critical importance to the success of missing feature methods. Conventional mask estimation methods (e.g. Vizinho et al., 1999) rely on noise-estimation techniques, such as those used in spectral subtraction (Boll, 1979) to estimate the local SNR at each time-frequency location. Each element's SNR estimate is compared to a specified threshold and the element is labelled accordingly. Such mask estimation methods perform well when the corrupting noise is stationary or slowly-varying, as this assumption is required for spectral subtraction. However, when the noise is non-stationary, masks estimated in this manner can be very inaccurate. In Fig. 1, cluster-based reconstruction (Raj et al., 2000), a feature-compensation missing feature method, has been applied to noisy speech using masks estimated via spectral subtraction and "oracle" masks generated from full a priori knowledge of the noise signal. The figure on the left 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. The figure on

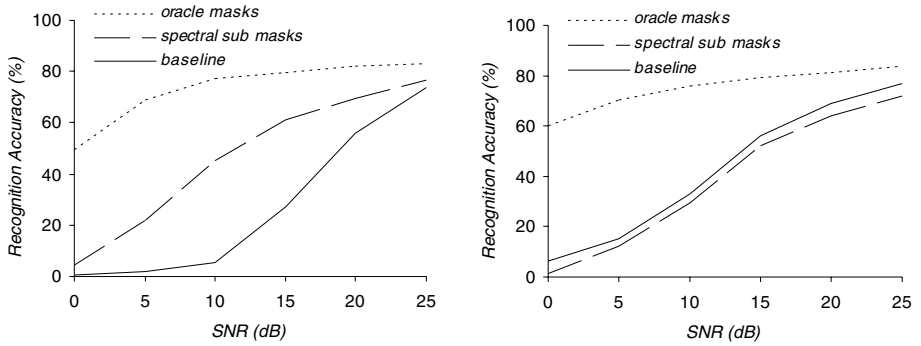

Fig. 1. Recognition accuracy vs. SNR when missing feature methods are applied to speech that has been corrupted by white noise (left) and speech that has been corrupted by music (right).

the right shows the same plot for speech corrupted by music, which is highly non-stationary. Here, spectral subtraction-based mask estimation completely fails. 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 show the potential of missing feature methods for noise compensation if the masks can be estimated reliably.

The mask estimation problem is further magnified when we consider the potential for larger improvements in recognition accuracy that has been demonstrated using missing feature methods based on "soft" mask decisions. Simple noise-estimation techniques do not provide the probabilistic measurement of an element's reliability that these methods require. Two current methods of solving this problem have been proposed in the literature. The first, by Barker et al. (2000) attempts to convert a boundless spectral-subtraction-based SNR estimate into a bounded [0,1] measurement via a warping function such as a sigmoid. This method has been shown to be successful on speech corrupted with pseudo-stationary noises. However, because it relies on spectral-subtraction-based SNR estimates, it is still subject to the same stationarity assumption that hindered the spectral-subtraction-based binary mask estimation methods. It has the potentially more significant drawback that the mask values are not actually probabilities. They are SNR estimates presented on a different scale. The second soft-decision mask estimation method, by Renevey and Drygajlo (2001), overcomes this drawback and does compute actual probabilities. In this method, the noise corrupting the speech signal is assumed to follow a Gaussian distribution and the parameters of the distribution are estimated using the non-speech segments of an utterance. Using these estimated parameters, the probability that the signal-plus-noise to noise ratio (referred to as *a posteriori* SNR in (Renevey and Drygajlo, 2001)) is above a specified threshold is computed and used as the mask value. However, this method, while capable of computing true probabilities, still assumes that the corrupting noise is both stationary and Gaussian. Yet, there are many real-world noises where both of these assumptions are invalid.

In this paper we present a mask-estimation technique that uses a Bayesian classification strategy to determine the reliability of each spectrographic element (Seltzer et al., 2000). Classification is performed using a set of features representative of the characteristics of speech, with no explicit reference to the noise. Casting mask estimation as a Bayesian classification problem has four distinct advantages. First, the problem of mask estimation is reduced from the difficult task of noise or SNR estimation to a simpler classification task. Second, the classification scheme allows any information that is pertinent to be easily incorporated as features into the mask estimation decision process. Third, with an appropriate choice of features, mask estimation can be free of assumptions about the corrupting noise. Finally,

by using a Bayesian classification scheme, we are easily able to generate truly probabilistic spectrographic masks.

In Section 2 we describe the feature set used by the classifier-based mask estimator to estimate the spectrographic masks. In Section 3 we describe the classification strategy we use. Section 4 briefly describes the missing feature methods that we will use to test the mask estimation method proposed. We describe experiments that were performed to evaluate the proposed method in Section 5. Finally, we summarize our results and highlight directions for future research in Section 6.

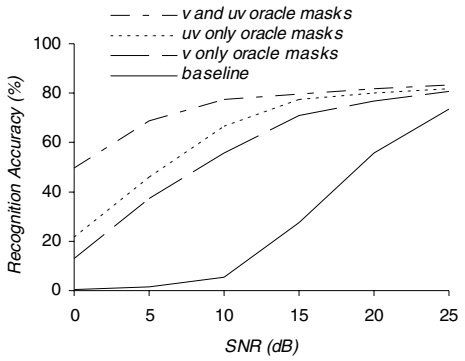

Fig. 2. Recognition accuracy vs. SNR on speech that has been corrupted by white noise. Missing feature compensation was applied using oracle masks applied to both the voiced and unvoiced segments, only the unvoiced segments, and only the voiced segments.

2. Feature extraction

As speech recognition systems become deployed in more areas, the variety of environments encountered and the types of noises which may corrupt the speech signal increase. We have learned that for robust speech recognition in unknown environments, it is preferable to extract cues directly from the speech signal, rather than try to estimate characteristics of the corrupting noise (Singh et al., 2001). This choice is quite intuitive, because in any given real-world environment, the variety of noise sources is virtually limitless, while the speech signal remains essentially invariant. It is much easier to study and model speech than to study and model every type of noise. With this in mind, we seek a mask estimation classifier which utilizes features designed to exploit the inherent characteristics of the speech signal itself while making few, if any, assumptions about the environmental noise. In focusing on speech properties, it is apparent that 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 estimate mask values for voiced and unvoiced segments of speech.

This distinction between voiced and unvoiced speech is important not just in the design of features for mask estimation, but in the performance of missing feature compensation algorithms themselves. As an example, Fig. 2 shows a plot of SNR vs. recognition accuracy for speech corrupted with white noise using the cluster-based reconstruction method (Raj et al., 2000). The uppermost curve shows the performance using full oracle masks. The dotted curve shows the recognition accuracy when only the unvoiced segments are compensated, and the voiced segments are left uncompensated. Similarly, the dashed curve shows the recognition accuracy achieved when only the voiced segments are compensated and the unvoiced segments are left untouched. As the figure clearly indicates, effectively compensating for the unvoiced segments is more important to recognition performance than compensating the voiced regions. This make intuitive sense when we consider that unvoiced speech is typically of lower energy than voiced speech. For a given global SNR, the unvoiced segments will be more corrupt than the voiced segments.

Yet, for a variety of reasons, the estimation of spectrographic masks for unvoiced speech segments is more difficult than for voiced segments. For instance, since unvoiced segments are relatively low in energy compared to voiced segments, they frequently end up as "negative energy" regions in spectral-subtraction-based noise-estimation schemes, because the energy of the noisy speech component is less than the estimate of the noise energy alone. This results in an erroneous SNR estimation which may result in mask error.

These observations further suggest that designing a mask estimation scheme that is not based

on noise-estimation techniques will result in significant improvements in performance over conventional mask estimation methods. The following sections describe the features designed for the mask estimation classifier for voiced and unvoiced segments of speech.

2.1. Features for voiced speech segments

Voiced speech is characterized largely by its strong periodicity and harmonicity arising from the strong presence of a fundamental frequency, or pitch, and its harmonics. Additionally, voiced speech has a definite spectral contour across frequency, with more energy present in the low frequencies, and tapering off at the higher frequencies. We attempt to exploit these characteristics with the several classification features. Of course, in order to utilize features that exploit the periodicity and harmonicity of voiced speech, we require a pitch estimator that is robust to noise. In this work we use the pitch estimator "get_f0" based on the RAPT algorithm (Talkin, 1995) and provided in the Entropic *xwaves* package. In the algorithm, initial pitch estimates are made using a normalized autocorrelation function and the final values are chosen from potential candidates via dynamic programming. Voiced/unvoiced decisions are also made by the pitch estimator.

2.1.1. Comb filter ratio

Because of the harmonic nature of voiced speech, the majority of the energy of a clean voiced speech signal resides in its harmonics (Morgan et al., 1997). 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.

A comb filter is constructed based on the pitch estimates to capture the energy present in the harmonics of voiced speech. We use an IIR comb filter implementation given by the following transfer function:

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

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. 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

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

If we assume that most energy in 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 plus 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 filter ratio (CFR). The CFR is given by

$$\text{CFR}[i,\omega] = 10\log_{10}\left(\frac{\sum_{n_i} y_{\text{comb}}[n_i,\omega]^2}{\sum_{n_i} y_{\text{combshift}}[n_i,\omega]^2}\right) \tag{3}$$

where y_{comb} and $y_{\text{combshift}}$ are the outputs obtained after the speech signal in frame i and subband ω has been passed through the comb and shifted comb filters, respectively.

Fig. 3 shows a plot of the average CFR for all voiced frames of an utterance of speech corrupted by music and by white noise at various SNRs. As the figure shows, the CFR is a reliable predictor of noise-level in the signal. Additionally, the lines show similar trajectories even though the corrupting noises are significantly different.

From an auditory scene analysis point of view, comb filtering can be interpreted as means of identifying harmonicity cues. Alternatively, the authors in (Barker et al., 2001) used an autocorrelelogram to identify these cues in order to generate a "harmonicity mask". Using the harmonicity mask in conjunction with a conventional SNR-based mask

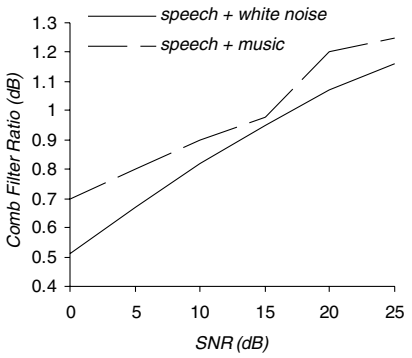

Fig. 3. Average comb filter ratio (CFR) vs. SNR for all voiced frames of an utterance corrupted by white noise and by music.

produced improved recognition results over using solely the SNR-based mask.

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, decreasing the ratio of the height of the largest secondary peak to the height of the main peak. We use this ratio 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. In general, 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 changes as a function of the spectral characteristics of the noise. We measure this impact as the log ratio of the energy in a subband to the overall frame energy as a measure of effect of additive noise on a particular subband and on the overall contour.

2.1.4. Kurtosis

Higher order spectra are used to capture information about a signal's deviation from Gaussianity (Donoho, 1981). Many real world audio signals, including speech, are generally regarded as a super-Gaussian signals; that is, their distribution has greater kurtosis than a Gaussian signal, with a sharper peak and more mass in the tails. When two super-Gaussian signals are combined, the kurtosis of the resulting signal typically goes down (Leblanc and De Leon, 1998). This characteristic has been exploited in the blind-source separation and speech enhancement literature (e.g. Leblanc and De Leon, 1998; Gillespie et al., 2001) where algorithms have been designed using kurtosis maximization objective functions. We assume as well that a clean speech signal and its noisy counterpart will have different kurtoses, and that we can capture this difference as a feature for classification. We use the kurtosis defined in (4), where expectations are estimated from sample averages in each subband of each frame.

$$\kappa_x = \frac{E\{x^4\}}{\{E\{x^2\}\}^2} - 3 \tag{4}$$

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.

2.1.6. 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. An accurate measure of the noise floor is difficult to obtain. We can, however, 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 segments and one at a higher energy representing high energy speech segments. The idea of statistically modeling the energy distributions of speech has been used for speech endpoint detection using HMMs (Acero et al., 1993). We have used a much simpler technique based on the noise-estimation technique in (Hirsch and Ehrlicher, 1995) 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 histogram 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. However, because this is one of many features in our classifier, we do not expect this to adversely affect the performance of the classifier in this situation.

2.1.7. Spectral-subtraction-based SNR estimate

As mentioned in Section 1, relying on assumptions of noise stationarity for mask estimation can result in poor recognition if the noise environment changes rapidly. However, in some environments, such a stationarity assumption is at least partially valid. For example, in an automobile or factory, the environmental noise can often be broken down into a stationary background noise and non-stationary impulses or other phenomena. In these quasi-stationary environments, spectrographic masks based on SNR estimation are able to provide improvement over the baseline recognition result (Vizinho et al., 1999). By including it as one of several features in our classifier, the SNR estimate can influence *but not exclusively control* the classification decision.

2.2. Features for unvoiced speech

Unvoiced speech is much more difficult to characterize than voiced speech. There is no harmonicity or other distinctive 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 less 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. The features that do not rely on pitch characterize a frame of speech in terms of the relative energy levels in each of the subbands, spectral shape, and statistical properties. They are useful features because we know that adding noise to a speech signal alters these characteristics. This is true for both voiced and unvoiced speech. For example, 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, all pitch-independent features used for classifying voiced speech can also be used for unvoiced speech. In our work, the mask estimation for the unvoiced segments is therefore performed using the features described above, with the exception of the two pitch-dependent features, the comb filter ratio and the autocorrelation peak ratio.

3. Classification strategy

To estimate the reliability of the spectrographic elements using the features described in the previous section, a two-class Bayesian classifier was designed. Each class, reliable and corrupt, was represented by a mixture of Gaussians with a single full-covariance matrix tied across all densities in the mixture. Because the number of features used for voiced and unvoiced speech differ (with unvoiced speech having two fewer features) we constructed a separate classifier for each type of speech. In addition, the values of each feature can vary significantly across frequency so a classifier was constructed for each subband as well.

Missing feature methods use an empirically-derived, method-dependent SNR threshold to determine whether a spectral component is reliable or corrupt (or in the soft-decision case, the degree of reliability or corruption). This threshold is used to label the data used to train the classifier.

The prior probabilities of the two-classes can be estimated from training data as the fraction of spectrographic components in each subband whose local SNR is above or below a given threshold. For an SNR threshold of 0 dB, Fig. 4a shows the prior probabilities of being reliable for both voiced and unvoiced speech, for each of the components of a log spectral vector computed using twenty Mel filters. Fig. 4b shows the average energy in each of the Mel subbands for both voiced and unvoiced speech, also computed from training data. Comparing the two figures, a clear correlation can be seen between the variation of the probability of being reliable across frequency and the energy profile of voiced and unvoiced speech across frequency. However, different missing feature methods operate differently and the effect of mask misclassifications on their performance is also different. Therefore, the priors estimated from the training data may not result in the best recognition. Improved performance can be obtained by tuning the prior probabilities using a cross-validation set.

4. Missing feature compensation methods

In missing feature compensation methods, the mask estimation and the missing feature algorithm itself work together to form a complete missing feature compensation system. To test the quality of the masks generated by the proposed Bayesian classification method, it is not enough to simply test mask classification accuracy against some ground truth such as oracle masks, constructed with full a priori knowledge of the SNR. Because there are many stages of processing between a missing feature algorithm and the final hypothesis output by the recognizer, measuring performance in terms of hits, misses, false alarms, and correct rejections, relative to the oracle mask, will not necessarily be a good indicator how the well the masks will perform with a given missing feature algorithm. Thus, the only "proper" measure of mask estimation performance is the recognition accuracy achieved when the estimated masks are used in conjunction with missing feature compensation methods. In fact, in soft-decision missing feature methods, there is no notion of ground truth for the masks, as all spectrographic elements are tagged in a probabilistic manner. Recognition accuracy is the *only* way to evaluate these masks.

We have tested the masks generated using two missing feature compensation methods. The first method, *cluster-based reconstruction* is a

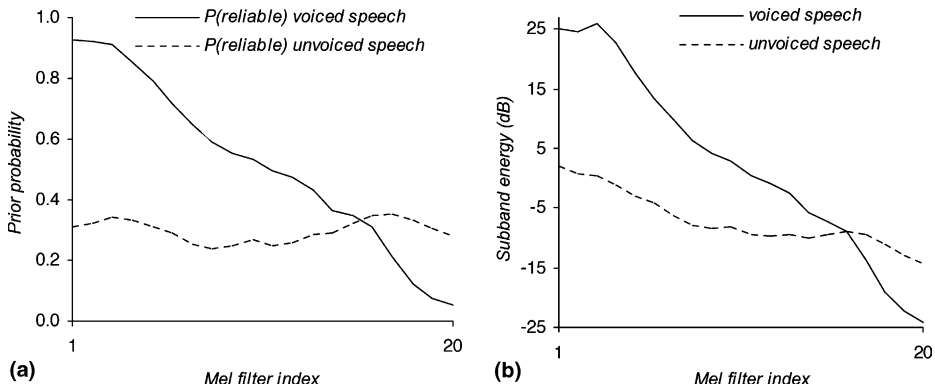

Fig. 4. (a) Prior probabilities of reliability of voiced and unvoiced speech from the mask estimation training data as a function of Mel filter index. The threshold for reliability is 0 dB SNR. (b) Average energy of voiced and unvoiced speech in each Mel subband as a function of Mel filter index.

feature-compensation method; all the processing takes place on the features, prior to recognition (Raj et al., 2000). It is a hard-decision method, so it requires spectrographic masks which label elements in a binary fashion. The second method, *soft-decision bounded integration* (also commonly called *soft-decision bounded marginalization*) is a classifier-compensation method; the incomplete log spectral vectors are used directly for recognition, and the basic manner in which class likelihoods are computed is modified inside the recognizer (Barker et al., 2000). As its name suggests, this is a soft-decision method. It requires probabilistic spectrographic masks which label the reliability of each element along a continuum of values between 0 and 1. A brief description of each method follows.

4.1. Cluster-based reconstruction

In the cluster-based reconstruction method, the log-spectral vectors of a training corpus of clean speech are grouped into a number of clusters using conventional expectation-maximization techniques (Dempster et al., 1977). The distributions of the vectors within each cluster are assumed to be Gaussian, and the mean, covariance, and a priori probability of each cluster are estimated from the training data. To compensate for noisy speech, the missing features are estimated by first identifying the cluster to which each corrupted log-spectral vector belongs, and then using the distributions of these clusters to estimate the missing elements of the vector. Cluster membership is given by the cluster k that has the highest likelihood of generating the noisy vector $S(t)$.

$$k_{S(t)} = \arg\max_k \{P(S(t)|k)P(k)\} \tag{5}$$

However, because $S(t)$ has unreliable elements, cluster membership cannot be identified in this way. The unreliable elements must first be integrated out of the cluster distributions so that cluster membership is estimated only from the components in vector that are present. Because the observed value (considered noisy or corrupt) represents the combined energy of the speech *and* the additive noise, we can use it as an upper bound for integration. Cluster membership is now given by (6) where $S_m(t)$ is a vector of the missing elements of vector $S(t)$ and $Y_m(t)$ is the vector of their observed values.

$$\hat{k}_{S(t)} = \arg\max_k \left\{ P(k) \int_{-\infty}^{Y_m(t)} P(S(t)|k)\, \mathrm{d}S_m(t) \right\} \tag{6}$$

Once the cluster membership k of a vector has been determined, missing feature reconstruction is performed using bounded MAP estimates based on the Gaussian distribution of the appropriate cluster and the upper bounds given by the observed corrupt values, as shown in (7).

$$\hat{S}_m(t) = \arg\max_{S_m} \Big\{ P(S_m(t)|S_0(t), \mu_{\hat{k}_{S(t)}}, \Sigma_{\hat{k}_{S(t)}}, S_m(t) \leqslant Y_m(t)) \Big\} \tag{7}$$

After the missing features have been reconstructed, cepstral coefficients can be extracted from the now-complete log spectral vectors in the usual manner and passed to a conventional HMM recognizer for decoding.

4.2. Soft-decision bounded integration

In general, bounded integration missing feature methods operate by marginalizing the components of the feature vector labelled as unreliable out of the HMM state distributions, again using the known value of the noisy element as an upper bound on the value of the component. The likelihood of each class is then computed using only the remaining "reliable" elements in the vector. This is considered bounded integration using hard decisions (Cooke et al., 2001). This method has recently been extended in (Barker et al., 2000) to use soft decisions on the element reliability. Under the soft-decision framework, each element $Y_i(t)$ is assigned a probability α that it is reliable and dominated by speech rather than noise. Likewise, each of these elements is assigned a probability $(1-\alpha)$ that the element is noise-dominated. Assuming all components in the vector are independent, the total likelihood of each component then becomes a weighted sum of the likelihood of the component

and its normalized cumulative probability. Mathematically, this can be expressed as

$$P(Y(t)|C) = \prod_i \left(\alpha(P(Y_i(t)|C)) + (1-\alpha)\frac{1}{Y_i(t)} \int_0^{Y_{i(t)}} P(S_i(t)|C)\,\mathrm{d}S_i \right) \quad (8)$$

In practice, state output densities are modelled by Gaussian mixtures with diagonal covariance matrices, and vector components are only assumed independent conditional on the Gaussian identity. In this case, (8) is applied to the individual Gaussians within the mixture.

5. Experimental results

Experiments in classifier-based mask estimation were performed using the DARPA Resource Management (RM) corpus (Price et al., 1988), 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, and in places, highly harmonic. The experimental procedures described in the following paragraphs were followed for each of the three noise environments.

Noise-corrupted speech was passed through a Mel filterbank consisting of twenty triangular FIR filters. For each frame and each subband, the features for mask estimation were extracted along with the log spectra for missing feature compensation and recognition.

The mask estimation classifier was trained on 2880 utterances from RM, corrupted by noise to various SNRs. For each frame and each subband, the features were extracted as described in the Section 2. For training, the pitch estimates required for the pitch-dependent features were obtained from clean speech using the *xwaves* "get_f0" pitch estimation package. The local SNR was computed for every time-frequency location, and the training data were labelled by comparing the SNR to a threshold. For the cluster-based reconstruction, the SNR threshold for reliability was −5 dB (Raj, 2000), while the soft-decision bounded integration method has a threshold of 0 dB (Barker et al., 2001). For each subband and each type of speech (voiced or unvoiced), mixtures of three Gaussians were estimated using conventional EM. A global full-covariance matrix was also estimated for each mixture.

A cross-validation data set of 200 utterances from RM was used to estimate the prior probabilities of reliability. A single prior probability estimate was used across all SNRs and subbands for the mask estimation classifiers for both the voiced and unvoiced speech segments. Although this is believed to be sub-optimal, performing a comprehensive search for the best set of prior probabilities for each subband was considered to be too computationally costly. There was no overlap between the cross-validation, training, and test sets.

The test set consisted of 400 utterances from RM. The pitch estimates for the test set were derived directly from the noisy speech for all environments. Frames with a non-zero pitch estimate were classified using the voiced speech mask estimation classifier for the appropriate subband, while those with no pitch estimate were classified using the unvoiced mask classifiers. For mask estimation purposes, no distinction was made between silence segments and unvoiced speech, nor was any segmentation performed. For each utterance the posterior probability of reliability was estimated for every spectrographic location to generate a probabilistic spectrographic mask. For the cluster-based missing feature approach, the probabilistic mask was converted to a binary mask using a probability of 0.5 as the threshold.

The spectrographic masks were estimated from the noise-corrupted test data. For comparison, masks were also estimated using the conventional noise-estimation-based techniques described in Section 1. Binary masks were estimated using the spectral subtraction approach with a threshold of 2.5 dB. This threshold was empirically determined to be optimal for spectral-subtraction-based masks applied to cluster-based reconstruction in (Raj, 2000). For the soft-decision bounded integration method, soft masks were developed using spectral-subtraction-based noise estimates using a

threshold of 0 dB in conjunction with a sigmoid warping function, as described in (Barker et al., 2000; Barker et al., 2001). A sigmoid is described by two parameters α and β. We initially set the sigmoid parameters to $\alpha = 3.0$, $\beta = 0.0$ which the authors reported in (Barker et al., 2001) to be optimal, but we obtained better performance with $\alpha = 0.5$.

Examples of masks estimated using each of these techniques are shown in Fig. 5 for a small segment of a test utterance. The black pixels represent reliable regions and the white pixels represent corrupt regions. Compared to the masks estimated using the noise-estimation-based techniques, the Bayesian masks capture more of the reliable regions of speech. Some of the reliable regions missed by the binary masks are present as shades of gray in both soft-decision masks. However, these regions are a darker gray in the Bayes mask, correctly indicating a higher probability of being reliable.

Once estimated, the masks were then used by each of the missing feature techniques for compensation. For cluster-based reconstruction, the features labelled as missing by the mask were reconstructed. After reconstruction, the now-complete twenty-dimensional log spectral vectors were converted to thirteen-dimensional cepstra via a DCT. Recognition was performed using the SPHINX-3 speech recognition system (Placeway et al., 1997). The system was trained on clean speech using the same 2880 utterances used to train the mask estimation classifier. Context-dependent continuous density HMMs were trained with one Gaussian per state. No delta or delta–delta features were used.

For the soft-decision bounded integration experiments, the noise-corrupted log spectra and the estimated spectrographic masks were processed by a modified version of SPHINX-3. Since this method operates exclusively in the log spectral domain, the training set was used to train

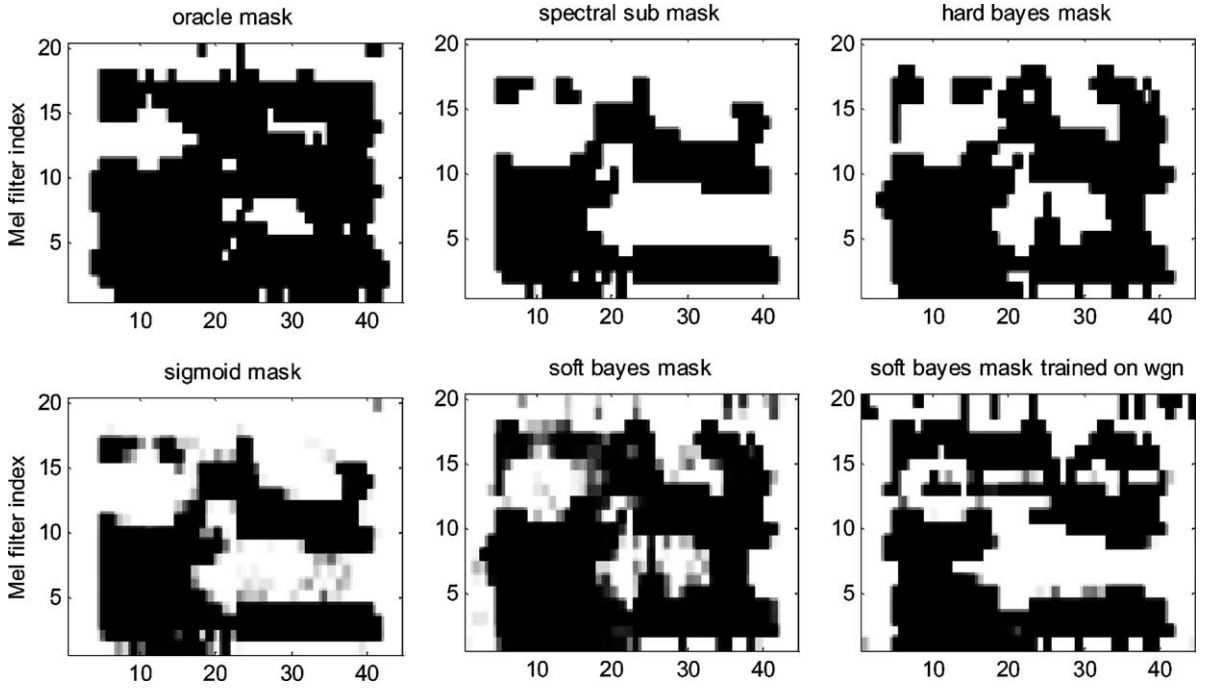

Fig. 5. Examples of spectrographic masks for a speech signal corrupted by factory noise to 10 dB SNR. The utterance is the word "EARLIER". Reliable regions are shown in black and corrupt regions are shown in white. The top row shows the oracle mask created from a priori knowledge of the local SNR, and estimated binary masks created by spectral subtraction and the proposed Bayes classifier method. The bottom row shows soft-decision masks estimated using the spectral-subtraction sigmoid method, the Bayes classifier method, and the Bayes classifier method trained on white noise.

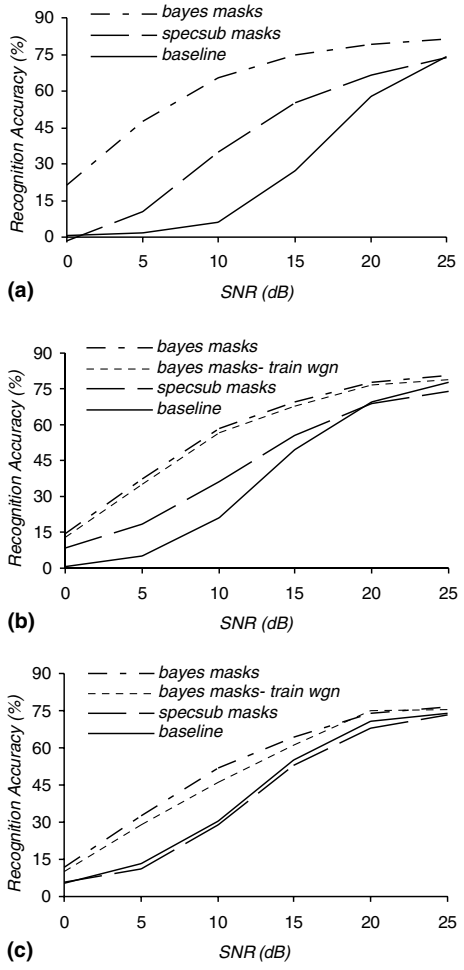

Fig. 6. Recognition accuracy vs. SNR using cluster-based feature reconstruction on speech corrupted by (a) white noise, (b) factory noise, and (c) music, with binary spectrographic masks estimated using the spectral-subtraction method, the proposed Bayes method, and in (b) and (c), the Bayes method using the mask classifier from (a) trained only on white noise.

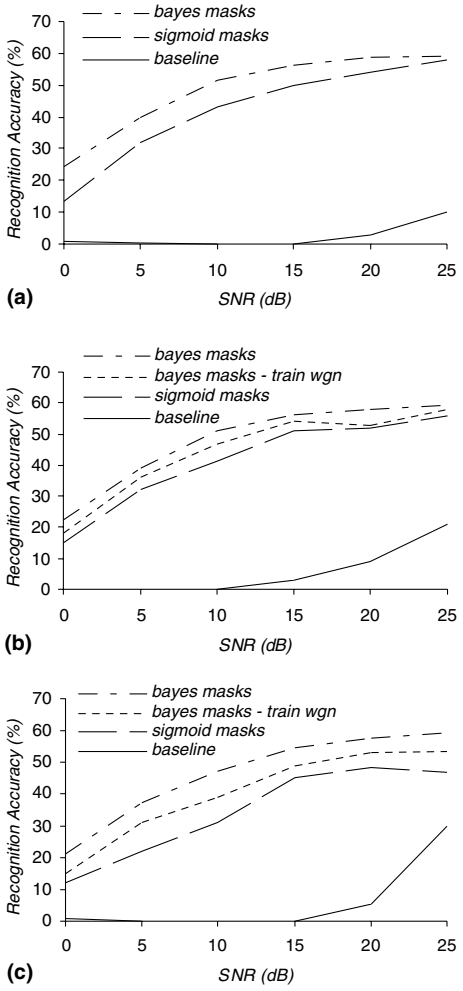

Fig. 7. Recognition accuracy vs. SNR using soft-decision bounded integration on speech corrupted by (a) white noise, (b) factory noise, and (c) music, with soft spectrographic masks estimated using the spectral-subtraction sigmoid method, the proposed Bayes method, and in (b, c), the Bayes method using the mask classifier from (a) trained only on white noise.

context-dependent continuous density HMMs on twenty-dimensional log spectra. Again, one Gaussian per state was used, and no delta or delta–delta features were used.

The recognition accuracies obtained for the three noise conditions using estimated masks with cluster-based reconstruction and soft-decision bounded integration are shown in Figs. 6 and 7, respectively. Figs. 6a and 7a show the performance for speech corrupted with white noise, Figs. 6b and 7b, speech corrupted by factory noise and Figs. 6c and 7c, speech corrupted by music. As the plots indicate, significant improvements in recognition accuracy were achieved for both missing feature methods using the proposed Bayesian mask estimation method, as compared to the noise-estimation-based techniques.

In the previous series of experiments, the noise environment, though not the SNR, was assumed to be known a priori. In some situations, this is a

realistic assumption. However, there are many situations, e.g. mobile telephony, where this assumption is not valid. Furthermore, the authors of Barker et al. (2000) noted that the sigmoid-based mask estimation technique is capable of better performance if the sigmoid parameters are tuned to match the test conditions.

To both test the performance of our mask estimation scheme in unknown environments and provide a fairer comparison to the sigmoid-based masks in the soft-decision case, the masks for the factory and music noise conditions were re-estimated, using the mask-estimation classifier that had been trained on speech corrupted only by white noise. Now, both the SNR *and* the noise type are unknown to the mask estimator. The last mask in Fig. 5 has been estimated in this manner. It has correctly identified almost all of the major blocks labelled by the mask trained with a matched noise type. However, it seems to be making more "extreme" decisions, as there are fewer grey regions compared to its matched-environment counterpart.

The performance of environment-independent mask estimation is shown by the dotted curves in plots (b) and (c) of Figs. 6 and 7 for the feature-reconstruction method and the bounded integration method, respectively. As the plots in Fig. 6 indicate, the recognition accuracy obtained with the feature reconstruction method using classifier-based masks trained only on white noise is virtually identical to the accuracy obtained when the mask estimator has been trained on the matching noise type. For bounded integration, the recognition accuracy obtained using the environment-independent masks is somewhat lower than that achieved when the mask estimation classifier operates in a known environment. However, as Fig. 7 indicates, this performance is still better than that obtained by the noise-estimation-based masks.

In addition, classifier-based mask estimation was performed on clean speech, uncorrupted by additive noise. In this case, the mask estimator should, in principle, label all spectrographic components as "reliable". Although the mask estimator did mark several components as "corrupt" (predominantly in the silence regions), applying missing feature compensation to clean speech using the estimated mask resulted in no degradation in speech recognition accuracy.

6. Conclusions and future work

In this paper we have presented a new Bayesian classifier for spectrographic mask estimation for missing feature compensation. Our classifier-based technique operates on the principle that it is better to concentrate on information we can extract from the noisy signal about the underlying speech, rather than trying to estimate properties of the noise. We have shown significant and consistent improvements over conventional noise-estimation based mask techniques under a variety of noise types and SNRs using two different missing feature techniques, one requiring binary masks and one requiring probabilistic masks. Our classifier has been shown to operate successfully across multiple SNRs and noise types. We were also able to maintain good performance in situations where the environment is unknown to the mask estimator. This last result validates our goal of speech-focused features, rather than noise-focused features. When the noise changes, the features remain informative.

However, compared to previous methods, our classifier-based method is significantly more complicated. There are many more parameters to optimize. Indeed, we believe the current mask estimator is performing sub-optimally as a result of assumptions made for the sake of expediency. For example, we know that the prior probabilities are not equal across subbands and voiced and unvoiced speech. However, doing an exhaustive search of all combinations of priors is simply not feasible. Similarly, it may be the case that not all features perform equally well in all subbands. For example, the pitch-based features may be less informative in the higher subbands, since the energy in the higher harmonics is relatively low, and therefore, perhaps their contribution to the mask estimation decision should be de-weighted in those bands.

In addition, while Figs. 6 and 7 show our classifier-based mask estimation methods consistently improve over previous mask estimation methods

in a variety of environments, the plots of speech corrupted by music show the smallest relative improvement. In environments where both the speech and the corrupting noise are harmonic, accurate pitch estimation is a difficult problem, especially as low SNRs. As a result, the pitch-dependent mask estimation features may be unreliable. To improve the performance in these conditions, pitch estimation algorithms capable of processing multiple harmonic streams may be required.

Casting the mask estimation problem as one of Bayesian classification provides many opportunities to improve mask estimation performance. For example, we can apply unsupervised adaptation using MAP techniques or MLLR (Leggetter and Woodland, 1994) to improve mask estimation performance. Additionally, this work has demonstrated the importance of accurate mask estimation in the unvoiced speech segments. To improve performance in this area, we need to develop more informative features for unvoiced speech. These can easily then be incorporated into the classification scheme. In addition, the current classification scheme treats each spectrographic element independently. However, it is clear from observing the spectrographic masks that there is a high correlation between the reliability of neighboring pixels, indicated by the block-like nature of the reliable and corrupt regions. We plan to try to capture this information by incorporating context information around a given pixel into the classification framework, e.g. using a larger feature vector composed of an element's features and the features of its neighbors, and by exploring image processing techniques to post-process the mask estimates. Preliminary work in these areas was performed in (Seltzer, 2000).

Finally, the most challenging task remains developing a framework in which the missing feature methods and the speech recognizer are incorporated into the mask estimation procedure. This would allow estimated masks to be optimized directly for speech recognition performance, rather than mask estimation accuracy. Such an approach would provide the means for a more principled and ideally unsupervised search for optimal values of the parameters for mask estimation, which we believe will greatly improve the mask estimation performance.

Acknowledgments

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

Acero, A., Crespo, C., Torrecilla, J.C., 1993. Robust HMM-based endpoint detector. Proc. Eurospeech'93.

Barker, J., Josifovski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. Proc. ICSLP'00.

Barker, J., Cooke, M., Green, P., 2001. Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise. Proc. Eurospeech'01.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing 27 (2), 113–120.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication 34 (3), 267–285.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B 39, 1–38.

Donoho, D.L., 1981. On minimum entropy deconvolution. In: Findley, D.F. (Ed.), Applied Time Series Analysis. Academic Press, New York.

Gillespie, B., Malvar, H., Florencio, D., 2001. Speech dereverberation via maximum-kurtosis subband adaptive filtering. Proc. ICASSP'01.

Hirsch, H.G., Ehrlicher, C., 1995. Noise estimation techniques for robust speech recognition. Proc. ICASSP'95, pp. 153–156.

Leblanc, J.P., De Leon, P.L., 1998. Speech separation by kurtosis maximization. Proc ICASSP'98.

Leggetter, C. J., Woodland, P.C., 1994. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG/ TR. 181, Cambridge University, Engineering Department, Cambridge, June 1994.

Moreno, P.J., 1996. Speech recognition in noisy environments. Ph.D. dissertation, Carnegie Mellon University, May 1996.

Morgan, D.P., George, E.B., Lee, L.T., Kay, S.M., 1997. Cochannel speaker separation by harmonic enhancement and suppression. IEEE Transactions on Speech and Audio Processing 5 (5), 407–424.

Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., Stern, R., Thayer, E., 1997. The 1996 Hub-4 Sphinx-3 system. Proc. DARPA Speech Recognition Workshop.

Price, P., Fisher, W.M., Bernstein, J., Pallet, D.S., 1988. The DARPA 1000 word Resource Management database for continuous speech recognition. Proc. ICASSP'88, pp. 651–654.

Raj, B., 2000. Reconstruction of incomplete spectrograms for robust speech recognition. Ph.D. dissertation, Carnegie Mellon University, May 2000.

Raj, B., Singh, R., Stern, R.M., 1998. Inference of missing spectrographic features for robust speech recognition. Proc. ICSLP'98.

Raj, B., Seltzer, M.L., Stern, R.M., 2000. Reconstruction of damaged spectrographic regions for robust speech recognition. Proc. ICSLP'00.

Raj, B., Seltzer, M.L., Stern, R.M., 2001. Robust speech recognition: the case for restoring missing features. Proc. CRAC'01.

Renevey, P., 2000. Speech recognition in noisy conditions using missing feature approach. Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne.

Renevey, P., Drygajlo, A., 2001. Detection of reliable features for speech recognition in noisy conditions using a statistical criterion. Proc. CRAC'01.

Seltzer, M. L., 2000. Automatic detection of corrupt spectrographic features for robust speech recognition. Master's thesis. Carnegie Mellon University, May 2000.

Seltzer, M.L., Raj, B., Stern, R.M., 2000. Classifier-based mask estimation for missing feature methods of robust speech recognition. Proc. ICSLP'00.

Singh, R., Seltzer, M.L., Raj, B., Stern, R.M., 2001. Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination. Proc. ICASSP'01.

Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). In: Speech Coding and Synthesis. Elsevier Science, Amsterdam, NL, pp. 495–518.

Vizinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study. Proc. Eurospeech'99.