Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 52 (2010) 72–81

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Robust speech recognition by integrating speech separation and hypothesis testing

Soundararajan Srinivasan [a,*], DeLiang Wang [b,*]

[a] Biomedical Engineering Department, The Ohio State University, Columbus, OH 43210, USA

[b] Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210, USA

Received 27 October 2008; received in revised form 24 August 2009; accepted 24 August 2009

Abstract

Missing-data methods attempt to improve robust speech recognition by distinguishing between reliable and unreliable data in the time–frequency (*T–F*) domain. Such methods require a binary mask to label speech-dominant *T–F* regions of a noisy speech signal as reliable and the rest as unreliable. Current methods for computing the mask are based mainly on bottom-up cues such as harmonicity and produce labeling errors that degrade recognition performance. In this paper, we propose a two-stage recognition system that combines bottom-up and top-down cues in order to simultaneously improve both mask estimation and recognition accuracy. First, an *n*-best lattice consistent with a speech separation mask is generated. The lattice is then re-scored by expanding the mask using a model-based hypothesis test to determine the reliability of individual *T–F* units. Systematic evaluations of the proposed system show significant improvement in recognition performance compared to that using speech separation alone.

Keywords: Robust speech recognition; Missing-data recognizer; Ideal binary mask; Speech segregation; Top-down processing

1. Introduction

The performance of automatic speech recognizers (ASRs) degrades rapidly in the presence of noise and other distortions (Gong, 1995; Huang et al., 2001). Speech recognizers are typically trained on clean speech and face a problem of mismatch when used in conditions where speech occurs simultaneously with other sound sources. To mitigate the effect of this mismatch on recognition, noisy speech is typically preprocessed by speech enhancement algorithms (Loizou, 2007), such as spectral subtraction based systems (Boll, 1979; Droppo et al., 2002). If samples of the corrupting noise source are available *a priori*, a model for the noise source can additionally be trained and noisy speech may be jointly decoded using the trained models of speech and noise (Varga et al., 1990; Gales and Young, 2007) or enhanced using linear filtering methods (Ephraim, 1992). However, in many realistic applications, the performance of the above approaches to robust speech recognition is inadequate (Cooke et al., 2001).

To deal with the mismatch issue, a missing-data approach to robust speech recognition has been proposed by Cooke et al. (2001). This method distinguishes between reliable and unreliable data in the time–frequency (*T–F*) domain. When speech is contaminated by additive noise, some *T–F* regions will contain predominantly speech energy (reliable) and the rest are dominated by noise energy. The missing-data ASR treats the latter *T–F* units as missing or unreliable during recognition. The missing-data recognizer, therefore, requires a binary *T–F* mask that provides information about which *T–F* units are reliable and which are unreliable. Previous studies have shown that the missing-data recognizer performs very well

[*] Corresponding authors. Present address: Robert Bosch LLC, Research and Technology Center North America, Pittsburgh, PA 15212, USA (S. Srinivasan). Tel.: +1 412 325 8452 (S. Srinivasan), +1 614 292 6827 (D.L. Wang).

E-mail addresses: srinivasan.36@osu.edu, soundar.srinivasan@us.bosch.com (S. Srinivasan), dwang@cse.ohio-state.edu (D. Wang).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.08.008

when this mask is known *a priori* (Cooke et al., 2001; Roman et al., 2003; Barker et al., 2005; Srinivasan et al., 2006). Attempts to estimate such a binary mask through front-end preprocessing using speech separation techniques have been only partly successful. Spectral subtraction is frequently used to generate such binary masks in missing-data studies (Drygajlo and El-Maliki, 1998; Cooke et al., 2001). For this purpose, noise is usually assumed to be long-term stationary and its spectrum is estimated from frames that do not contain speech (speech silent frames containing just background noise). The noise spectrum is then used to estimate the SNR in each *T–F* unit. If the SNR in a *T–F* unit exceeds a threshold, it is labeled reliable; it is labeled unreliable otherwise. In the presence of non-stationary interference sources, however, the use of spectral subtraction results in a poor estimate of the mask. Methods that primarily utilize the harmonicity of voiced speech have also been proposed to estimate the mask for missing-data applications (Seltzer et al., 2000; Brown et al., 2001; van Hamme, 2004). However, these methods are unable to deal with unvoiced speech. Accurate estimation of pitch is also difficult, if not impossible, when the SNR is low. Hence, the estimated binary mask corresponding to voiced speech may not be reliable. Therefore, good estimation of the binary *T–F* mask remains a challenging problem.

On the other hand, the human auditory system exhibits a remarkable ability to segregate a target speech source from various interferences (Darwin, 2008). According to Bregman (Bregman, 1990), this is accomplished via a process termed auditory scene analysis (ASA). ASA involves two types of organization, primitive and schema-driven. Primitive ASA is based on bottom-up cues such as pitch and spatial location of a sound source. Schema-based ASA is based on top-down use of stored knowledge about auditory inputs, e.g. speech patterns, and supplements primitive analysis. Top-down information has also been used successfully in computational ASA studies previously (Barker et al., 2005; Srinivasan and Wang, 2005b). In particular, Barker et al. (2005) have proposed a top-down approach to identify *T–F* units that are dominated by speech in a noisy mixture. We believe that a top-down approach, using speech models, can be used to refine the mask generated by bottom-up processing to achieve improved recognition results.

In this paper, we present a two-pass missing-data recognition system that estimates an ideal binary *T–F* mask and improves recognition results at the same time. A *T–F* unit in the ideal binary mask is 1 if in the corresponding *T–F* unit the noisy speech contains more speech energy than interference energy; it is 0 otherwise. The ideal binary mask is obtained a priori from premixed speech and noise. In the first pass, a mask produced by a speech separation system is used to generate an *n*-best lattice using a missing-data recognizer. This corresponds to bottom-up processing. This lattice is then re-scored, to produce the final recognition results by augmenting the initial mask using the information contained in states along individual paths. Specifically, we propose a state-based hypothesis test to determine the reliability of each *T–F* unit. This corresponds to top-down analysis. The resulting recognition accuracy is substantially better than that of the conventional ASR as well as the missing-data recognizer using the mask produced by speech separation alone.

The rest of the paper is organized as follows. The next section contains a detailed presentation of the system. The proposed system has been systematically evaluated on a noisy connected digit recognition task and the evaluation results are presented in Section 3. Section 4 concludes the paper.

2. System description

The proposed system is a two-pass recognition system as shown in Fig. 1. In the first pass, we use an initial, conservative mask generated through bottom-up separation as input to a missing-data recognizer. The output of the missing-data ASR is a lattice containing *n*-best hypotheses. The initial mask is then augmented by another mask generated through spectral subtraction to result in a three-way mask. In the second pass, we use a state-based hypothesis test to refine this three-way mask and improve recognition results at the same time.

2.1. Bottom-up speech separation

The input to the system is a mixture of speech and interference, sampled at 20 kHz. Following the original study of Cooke et al. (2001), we use an auditory filterbank decomposition (Patterson et al., 1988) of the input signal to generate feature vectors for recognition. Specifically, the input is first analyzed using a 128 channel gammatone filterbank whose center frequencies are quasi-logarithmically spaced from 80 Hz to 5 kHz (see (Wang and Brown, 2006) Chapter 1). Our previous studies (Srinivasan, 2006) have shown that this frequency range is adequate for recognition of male speech considered in this study (see Section 3). The instantaneous Hilbert envelope at the output of each gammatone filter is then downsampled to a frame rate of 100 Hz and finally cube-root compressed (Cooke et al., 2001). As a result, the input signal is decomposed into a two-dimensional matrix of *T–F* units.

The missing-data recognizer (Cooke et al., 2001) makes use of spectro-temporal redundancy in speech to recognize a noisy signal based on its speech-dominant *T–F* units. Specifically, it modifies the computation of the observation probability in a state of an HMM-based ASR to handle missing or unreliable data. The observation density in a conventional ASR is typically modeled using a mixture of Gaussians as shown below:

$$p(x|q) = \sum_{k=1}^{M} p(k|q)p(x|k,q), \tag{1}$$

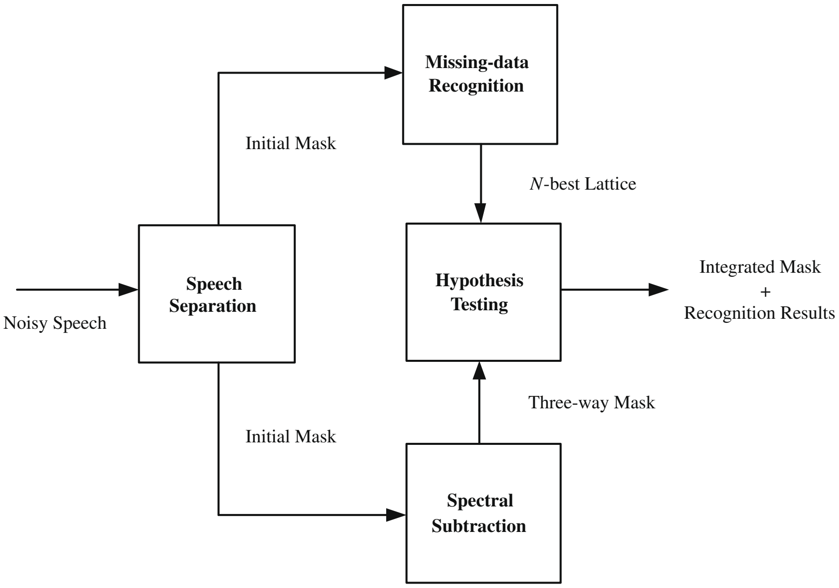

Fig. 1. Schematic diagram of the proposed two-pass recognition system. The noisy input is processed by a speech separation system to produce an initial binary *T–F* mask. This mask is used by the missing-data recognizer to generate an *n*-best lattice. The states along this lattice are used in a hypothesis test to refine a three-way mask generated by combining the initial mask with a mask produced by spectral subtraction. The system outputs an integrated mask along with the recognition results. Notice that the information flows bottom-up leading to the lattice generation and then top-down leading simultaneously to mask estimation and final recognition.

where x is the spectral energy feature vector in a frame, M is the number of mixture components, k is the component index, q is an HMM state, $p(k|q)$ is the component weight, and $p(X|k,q) = N(X;\mu_{k,q},\Sigma_{k,q})$. When parts of x are corrupted by noise, the missing-data ASR treats them as unreliable data (Cooke et al., 2001) and marginalizes those feature dimensions in the computation of the likelihood in (1). In most missing-data studies, the various dimensions of the feature vectors are modeled as independent given a mixture. Theoretically, this is a good approximation if an adequate number of mixture components are used (McLachlan and Basford, 1988). Hence, the computation of the observation density is modified in the presence of unreliable data as:

$$p(x|q) = \sum_{k=1}^{M} p(k|q) \prod p(x_{r,j}|k,q) \prod \int p(x_{u,i}|k,q)dx_{u,i}, \quad (2)$$

where $x_{r,j}$ and $x_{u,i}$ correspond to the spectral energies in a reliable (j) and an unreliable (i) feature dimension, respectively. Under additive noise conditions, the true speech value $\tilde{x}_{u,i}$, in the unreliable part may be constrained as $0 \leqslant \tilde{x}_{u,i} \leqslant y_{u,i}$ (Cooke et al., 2001), where $y_{u,i}$ is the observed (noisy) spectral energy. This constraint is then used as bounds on the integral in (2) as:

$$p(x|q) = \sum_{k=1}^{M} p(k|q) \prod p(x_{r,j}|k,q) \prod \int_0^{y_{u,i}} p(x_{u,i}|k,q)dx_{u,i}. \quad (3)$$

Note that the computation of $\int_0^{y_{u,i}} p(x_{u,i}|k,q)dx_{u,i}$ involves the calculation of two error functions (Stark and Woods, 2002) as shown below.

$$\int_0^{y_{u,i}} p(x_{u,i}|k,q)dx_{u,i} = erf\left(\frac{y_{u,i} - \mu_{k,q,i}}{\sigma_{k,q,i}}\right) - erf\left(\frac{0 - \mu_{k,q,i}}{\sigma_{k,q,i}}\right), \quad (4)$$

where $\mu_{k,q,i}$ and $\sigma_{k,q,i}$ are the mean and the standard deviation of the Gaussian density corresponding to the ith dimension.

As stated before, a fundamental requirement of the missing-data recognizer is a binary mask that informs it whether a *T–F* unit is reliable (value 1) or unreliable (value 0). This mask is usually generated through front-end processing such as those based on estimates of local SNR or those based on harmonicity of voiced speech. While the accuracy of the mask produced by such bottom-up speech separation methods is good in limited situations, it may contain a large number of errors under noisy conditions. Our experiments with the missing-data recognizer have shown that wrongly labeling unreliable *T–F* units as reliable is especially harmful for recognition. Hence, in the first pass, we use a "conservative" binary *T–F* mask as input to the missing-data recognizer. In most cases, the conservative mask may be obtained by simple modifications to existing bottom-up algorithms. For example, if the mask estimation is based on an estimate of the local SNR in a *T–F* unit, the SNR threshold above which the *T–F* unit is labeled 1 could be increased to produce a conservative mask.

Fig. 2 shows an example of a conservative binary mask produced using spectral subtraction. Fig. 2a shows the cochleagram of a speech signal, "NINE ZERO ZERO ONE NINE NINE EIGHT", from the TIDigits database (Leonard, 1984). A cochleagram is a *T–F* representation of a signal analogous to a spectrogram, but is generated

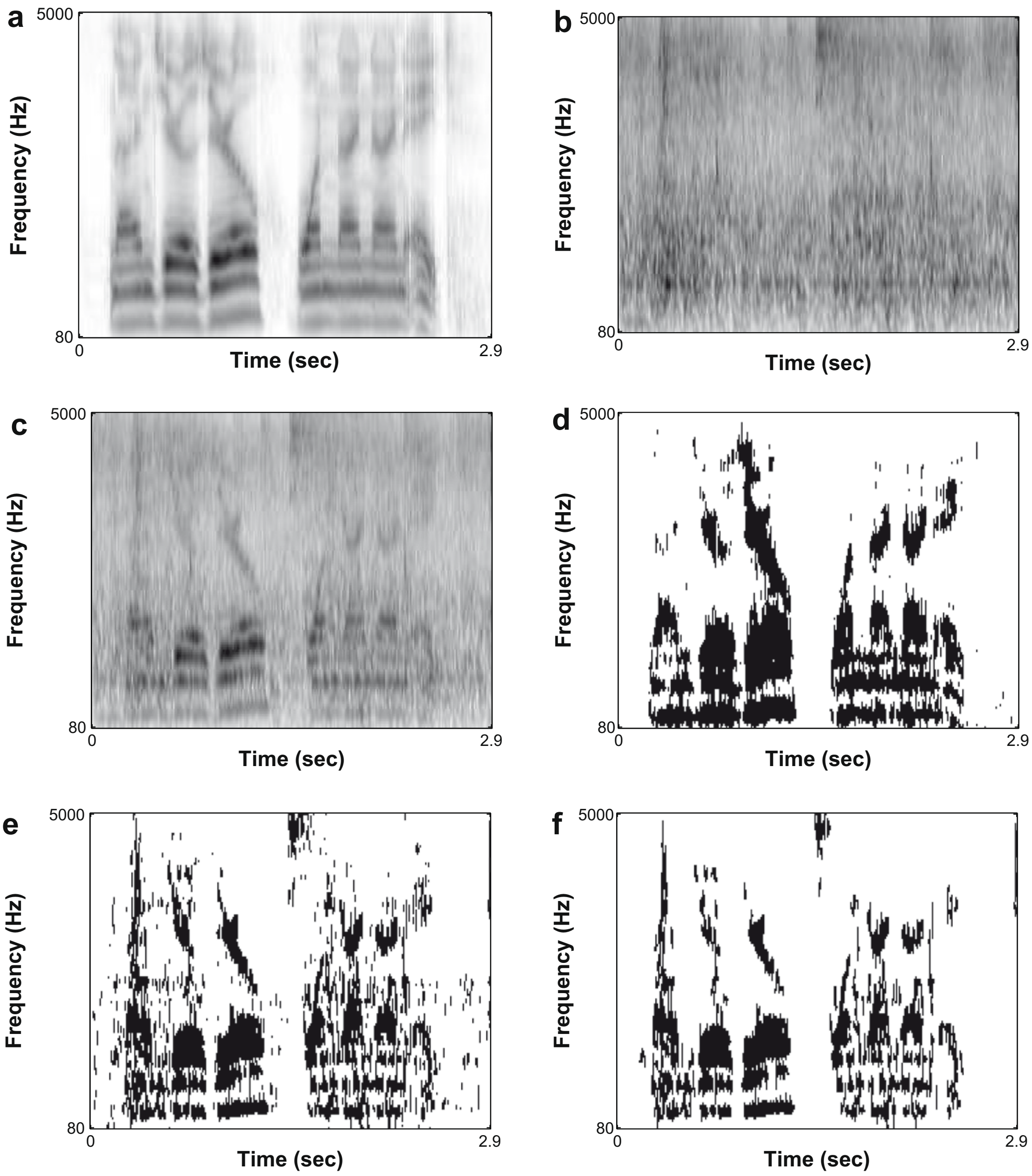

Fig. 2. An illustration of the conservative binary *T–F* mask generated using spectral subtraction. (a) The cochleagram of a speech signal. (b) The cochleagram of factory noise. (c) The cochleagram of the mixture. (d) The ideal binary mask. (e) The binary mask obtained from spectral subtraction using a local SNR threshold of 0 dB. (f) The binary mask obtained from spectral subtraction using a local SNR threshold of 7.7 dB. Speech-dominant *T–F* units in (d–f) are marked black and the noise-dominant ones white.

using the gammatone filterbank decomposition of a signal described before (see (Wang and Brown, 2006)). Fig. 2b shows the cochleagram of a factory noise source from the NOISEX corpus (Varga et al., 1992). The cochleagram of a mixture of the speech signal from Fig. 2a and the noise source from Fig. 2b at a SNR of 0 dB is shown in Fig. 2c. The ideal binary mask corresponding to this mixture is shown in Fig. 2d. *T–F* units labeled 1 in the mask are shown in black; the rest are shown in white. Fig. 2e shows the binary mask produced by thresholding *T–F* units based on the estimated local SNR. A threshold of 0 dB is used. Spectral subtraction is used in this figure to estimate the local SNR. Notice that the mask contains significant deviations from the ideal binary mask. In particular, in frames that do not contain any voice activity, the mask erroneously labels many *T–F* units as speech-dominant. Fig. 2f shows the mask produced using a threshold of 7.7 dB. This choice of a conservative threshold significantly reduces the wrong labeling of noise-dominant *T–F* units as reliable. Note that the conservative mask may be produced by using any other bottom-up speech separation system too. For example, a system using harmonicity of voiced speech could increase the threshold used in assessing the periodicity similarity in a *T–F* unit (see e.g. (Wang and Brown, 1999)).

A conservative mask is needed to ensure that the first pass retains a small set of viable recognition candidates for subsequent top-down analysis. This will not only reduces the search space during the second pass, but also ensures that the less likely recognition candidates are not considered in subsequent mask determination. This mask is used in conjunction with the spectral feature vectors as input to the missing-data ASR in the first pass. The output of this stage is a lattice from which an *n*-best hypothesis list can be generated.

2.2. Top-down hypothesis testing

In the second stage, we seek to augment the bottom-up mask by using top-down information in the form of speech models. Specifically, we use those states of the missing-data ASR that are contained in the *n*-best lattice.

The use of a conservative criterion during bottom-up mask generation ensures a high probability that a *T–F* unit labeled reliable is actually dominated by speech. Hence, during top-down processing, we only analyze the *T–F* units labeled 0 in the first stage. Since the number of unreliable *T–F* units under low SNR conditions is very high, a state-based analysis of each unreliable *T–F* unit is computationally prohibitive. Therefore, we only analyze those 0-labeled *T–F* units which have a high probability of being relabeled 1. For this purpose, we use spectral subtraction: The spectrum of noise is estimated as the average spectrum of the first 10 frames of the noisy speech spectrum (Cooke et al., 2001). The noise spectrum is then used to estimate the local SNR in each *T–F* unit. Each *T–F* unit labeled 0 by bottom-up separation is now relabeled using a threshold δ_2 as below,

$$label = \begin{cases} 2 & \text{if} \quad SNR_{local} \geqslant \delta_2, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The choice of δ_2 represents a trade-off between increasing the computation time of top-down analysis and possibly reducing the number of speech-dominant units. The optimal value of δ_2 is dependent on SNR (Renevey and Drygajlo, 2001). For simplicity, we set δ_2 to a constant value of 0 dB as suggested by Barker et al. (2005). Not all *T–F* units labeled 2 are dominated by speech due to the limitations of spectral subtraction stated previously. Therefore, top-down processing is needed to remove noise-dominant *T–F* units from those labeled 2. Note that *T–F* units labeled 1 by bottom-up processing are not affected by this spectral subtraction based labeling. As a result, we now have a three-way mask.

The lattice generated by the first pass is now re-scored using the missing-data recognizer and the three-way mask. During re-scoring, each active state independently analyzes the *T–F* units labeled 2. The observation density of each state in an HMM-based ASR models a particular class of speech signal. This information could therefore be used to verify whether the observed value in a *T–F* unit is consistent with a speech state. This corresponds to top-down processing.

More generally, the task of deciding whether a test sample is inconsistent with pretrained models is known as novelty or outlier detection (Markou and Singh, 2003). Bishop (1994) suggests that unseen test samples should be first modeled using a uniform distribution. Subsequently, a threshold on the likelihood ratio of the pretrained distribution and the uniform distribution could be used to label a particular test data as “novel”. If training data follows a Gaussian distribution, Tax and Duin (1998) suggest that test samples outside of three standard deviations from the mean should be treated as outliers. The mean and the standard deviation again pertain to training data. When HMMs are used to model the training data, the observed value may be used to generate “evidence” and “counter-evidence” measures for speech in a particular state (Barker et al., 2005). In this paper, we use two measures to construct a hypothesis test for each *T–F* unit labeled 2 as follows: If a *T–F* unit satisfies the inequality in (6), we label it as 1; it is labeled 0 otherwise.

$$\sum_{k=1}^{M} p(k|q)\frac{p(y_i|k,q)}{p(y_i)} > \sum_{k=1}^{M} p(k|q)\int_0^{y_i} p(x_i|k,q)\frac{1}{\alpha}dx, \tag{6}$$

where

$$\alpha = \int_0^{y_i} p(x_i)dx, \tag{7}$$

is the normalization factor. Recall that y_i and x_i are, respectively the observed spectral energy and the speech spectral energy seen during training in frequency channel *i*. $p(y_i|k,q)$ is the likelihood of observing y_i given a state q and a mixture component k. The prior $p(x_i)$, used to normalize the probabilities in (6), is modeled empirically using a mixture of Gaussians based on the data used in training the ASR. The left hand side of (6) therefore models the evidence for the speech signal in state q. The right hand side provides the probability that the observed value is speech corrupted by additive noise, or counter-evidence for clean speech. For example, a high value provides evidence against the observed spectral energy being from speech alone. As stated previously, since y_i represents spectral energy, under additive noise conditions the range for the true speech value is $[0, y_i]$. In the absence of any knowledge about the noise level, (6) represents a conservative decision. This is consistent with our observation regarding the consequences of wrongly labeling unreliable *T–F* units as reliable. Note that this top-down mask refinement is performed simultaneously with the lattice re-scoring. Thus, we first use bottom-up cues to generate a conservative mask and then subsequently augment this mask using top-down processing.

As mentioned before, Barker et al. (2005) have also proposed a top-down approach using an HMM-based speech recognizer to identify *T–F* units that are dominated by speech in a noisy mixture. The primary difference in our approach is the use of bottom-up cues in our initial mask generation. This mask helps to drastically reduce the search space for top-down analysis (see Section 4). Additionally, two other differences are worth pointing out. First, while the "evidence" and "counter-evidence" measure definitions in (6) are also used by Barker et al. (2005), we use (6) to construct a hypothesis test for each *T–F* unit labeled 2. Barker et al. (2005), on the other hand, use the two measures to create alternate decoding traces corresponding to target and interference. Second, the prior distribution in their probabilistic framework is modeled as an uniform distribution which can cause a bias toward labeling more *T–F* units as unreliable (Barker et al., 2005). By empirically modeling the prior density using Gaussian mixtures, our system does not suffer from this bias.

3. Evaluation results

We have evaluated the system on speaker-independent recognition of connected digits. This task is also used by Cooke et al. (2001) and Barker et al. (2005). Thirteen (1–9, silence, short pause between words, zero and oh) word-level models are trained. All except the short pause model have 8 states, whose output distribution is modeled as a mixture of 10 Gaussians (Cooke et al., 2001). The short pause model has three states. The TIDigits database's male speaker data (Leonard, 1984) is used for both training and testing. This subset is commonly used in robust speech recognition studies (see example, (McCowan and Bourlard, 2003; Roman et al., 2003)); we expect that the evaluation results will be similar when tested on the female speaker data. Specifically, the models are trained using 4235 utterances from 55 speakers in the training set of this database. Testing is performed on a subset of the testing set consisting of 461 utterances from 6 speakers that are different from the speakers in the training set. This test set size ensures that the evaluation results are statistically significant. The speech prior is modeled empirically using a mixture of 100 Gaussians using all utterances in the training set. An HMM toolkit, HTK (Young et al., 2000), is used for training. For testing, a decoder incorporating our mask generation and missing-data recognition is used. To test the robustness of the proposed two-pass system on the aforementioned task, noise is added at a range of SNRs from −5 dB to 10 dB in steps of 5 dB. The noise source is the factory noise from the NOISEX corpus (Varga et al., 1992). Factory noise is chosen as it has energy in the formant regions, therefore posing challenging problems during recognition. It is also impulsive, making it difficult to estimate its spectrum using spectral subtraction methods (Cooke et al., 2001).

Monaural CASA systems that estimate an ideal binary mask have been used as front-ends for the missing-data recognizer previously (e.g. (Brown et al., 2001)). To generate this mask, target and interfering signals are decomposed into a two-dimensional matrix of *T–F* units using the procedure described in Section 2.1. The resulting cube-root compressed envelope is used as the estimate of energy (Cooke et al., 2001).

While several systems estimate the ideal binary mask well in low frequencies, they perform poorly in high frequencies (for an exception, see (Hu and Wang, 2004)). Additionally, under noisy conditions, high-frequency components of speech are more corrupted than low-frequency ones. Hence, to reveal the potential for our top-down processing stage, we set high-frequency components (above 1 kHz) of the ideal binary mask to 0 and use it as our bottom-up, speech separation (initial) mask for the first pass. Note that 1 kHz is commonly used as the low-frequency boundary in several binary mask estimation systems (see e.g., (Hu and Wang, 2004)). Here we use 2 best tokens in each state to generate the *n*-best lattice. Our experiments indicate that the use of only 2 tokens in each state is sufficient to retain the true hypothesis most of the time, while maintaining a reasonable computational cost. The top-down processing stage is then used to identify reliable regions above 1 kHz. Table 1 summarizes the performance of the two-pass system when using the ideal binary mask below 1 kHz ("Two-pass Mask"). Performance is measured in terms of word-level recognition accuracy at various SNRs. For comparison, the performance of the missing-data recognizer when using the speech separation mask is also shown ("One-pass Mask"). Additionally, we show the performance of the missing-data recognizer when using the ideal binary mask at all frequencies ("Ideal Binary Mask"), which represents the ceiling performance for the proposed approach.

Across all SNRs, the proposed system shows significant performance improvement over that of the missing-data recognizer using the ideal binary mask below 1 kHz. This indicates that the top-down hypothesis testing stage of our system is able to correctly identify reliable *T–F* units in the high-frequency region. The results also confirm previous findings that the missing-data recognizer achieves high accuracy with the ideal binary mask.

Fig. 3 illustrates the top-down mask refinement. Fig. 3a shows the mask corresponding to the mixture in Fig. 2c produced by using the ideal binary mask in the low frequencies and spectral subtraction in the high frequencies as described above. *T–F* units labeled 0 in the mask are shown in white, while those labeled 1 are shown in black. The gray *T–F* units correspond to those labeled 2 (see Section 2.2). Fig. 3b shows the mask produced by the top-down hypothesis testing stage. For ease of comparison, the ideal binary mask from Fig. 2d is also shown in Fig. 3c. Notice that compared to the ideal mask in Fig. 3a, the mask in Fig. 3b contains significantly less labeling of noise-dominant *T–F* units as 1. Additionally, the

Table 1
Digit recognition accuracy (%) of the proposed system and the missing-data recognizer using the initial mask.

System	SNR (dB)				
	−5	0	5	10	∞
One-pass Mask	63.8	68.7	68.6	75.6	92.7
Two-pass Mask	71.9	75.2	76	81.7	92.7
Ideal binary Mask	86.2	86.7	89.4	90.2	92.7

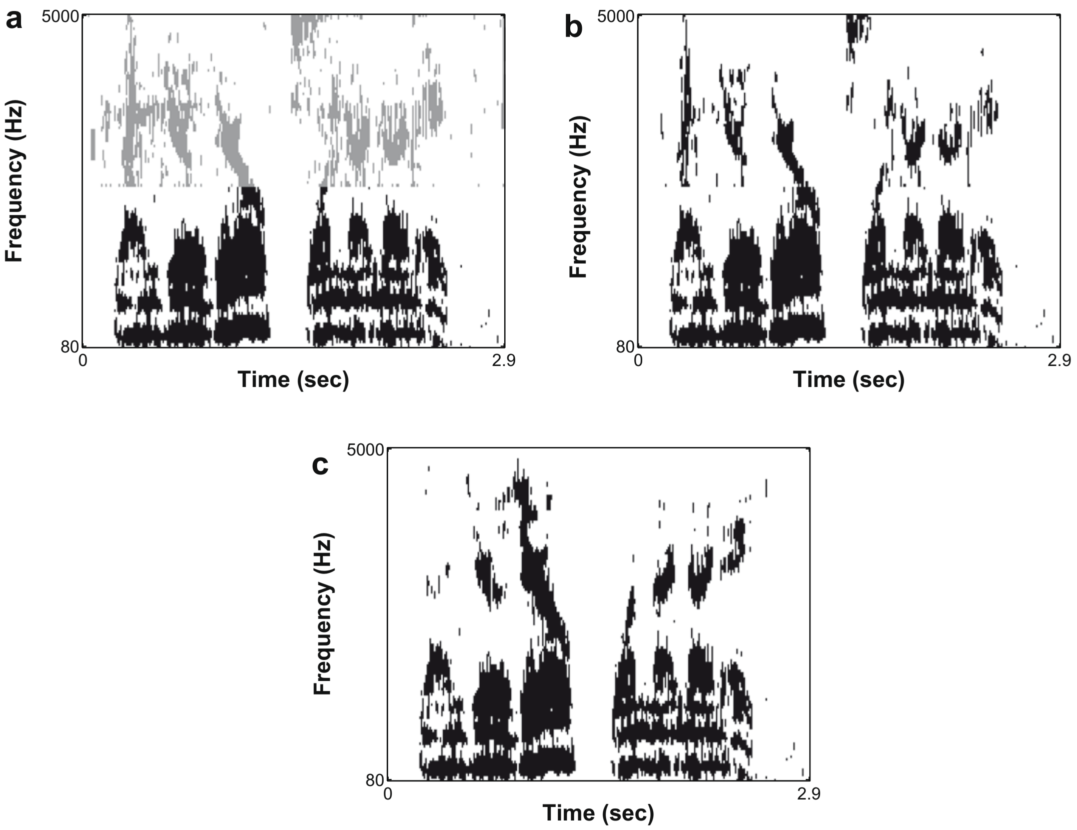

Fig. 3. An illustration of mask refinement by top-down analysis. (a) The three-way mask. The low-frequency *T–F* units in the mask correspond to the ideal binary mask. *T–F* units labeled 1 are marked black and the rest white. In high frequencies, spectral subtraction is used to label a *T–F* unit as 2 (gray) or 0 (white). (b) The mask produced by using the hypothesis test in (6). (c) The ideal binary mask.

mask in Fig. 3b is able to correctly retain all speech-dominant units from Fig. 3a.

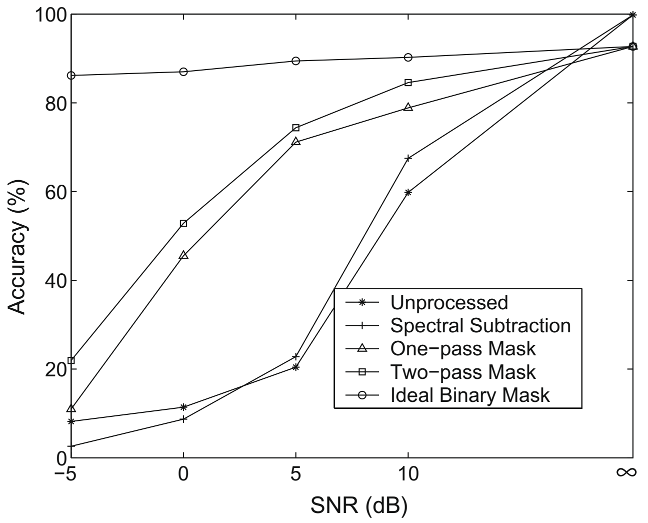

Fig. 4. Performance of the proposed system and the missing-data ASR using the mask produced by spectral subtraction. Two-pass Mask refers to the performance of the proposed system. Ideal Binary Mask and One-pass Mask refer to the performance of the missing-data recognizer using the ideal binary mask and mask from spectral subtraction, respectively. For comparison, the performance of the conventional ASR without the use of any front-end processing (Unprocessed) and with a spectral subtraction front-end processing (Spectral Subtraction) are also shown.

Next, we present results using two methods that estimate the ideal binary *T–F* mask in both high and low frequencies. First, we present results using spectral subtraction, which is a frequent choice for mask estimation in missing-data studies as mentioned in Section 1. Specifically, the noise spectrum is used to estimate the local SNR in each *T–F* unit as described in Section 2.2. Following the suggestion of Cooke et al. (2001), we use the threshold of 8 dB to generate our conservative, initial mask. Fig. 4 shows the performance of the proposed system using this mask for the first pass. Performance is shown in terms of digit recognition accuracy with respect to the SNR. For comparison, the performance of the missing-data ASR using the speech separation (spectral subtraction) mask, the ideal binary mask and that of a conventional MFCC-based ASR with no preprocessing and preprocessing using spectral subtraction are also shown. For spectral subtraction preprocessing, the spectrum of noise is once again estimated as the average spectrum of the first 10 frames of the noisy speech spectrum. The noise spectrum is then subtracted from the noisy speech spectrum to estimate the clean speech spectrum for processing by the conventional MFCC-based ASR. Across all SNR conditions, the proposed system shows significant improvement (p-value < 0.05) over the performance of the missing-data recognizer using the speech separation mask. An average reduction in word-error-rate (WER) of 16% is obtained. Note that performances of both the proposed system and the missing-data recognizer using the speech separation mask are substantially better than the unprocessed baseline. Also notice that for clean speech, the conventional MFCC-based ASR achieves the best performance; it is well known that recognition using cepstral coefficients yields superior performance compared to recognition using spectral coefficients under clean speech conditions (Davis and Mermelstein, 1980).

We now present results using a monaural CASA system that is able to handle both high and low frequencies of speech (Hu and Wang, 2004). This system is a voiced speech separation system based on two main stages: (1) segmentation, and (2) grouping. In segmentation, the input signal is decomposed into a collection of contiguous *T–F* segments (regions) that are dominated by one sound source. During grouping, those segments that are likely to belong to the same source are grouped together. In the low-frequency range, the system generates segments based on temporal continuity and cross-channel correlation, and groups them based on periodicity similarity. At high frequencies, the signal envelope fluctuates at the pitch rate and amplitude modulation rates are used for grouping (Hu and Wang, 2004). Provided a target pitch contour can be estimated, this segregation mechanism produces a binary mask which selects *T–F* units where speech dominates the interference. The system shows a robust performance when tested with a variety of noise intrusions. Hence, we use the system to generate our initial mask. For input to the system in Hu and Wang (2004), a pitch estimate is derived from the noisy mixture using Praat (Boersma and Weenink, 2002).

Fig. 5 summarizes the performance of the proposed system when using the mask generated by the Hu and Wang system as the initial mask. Performance is again shown in terms of digit recognition accuracy across different SNRs. Similar to the results obtained using spectral subtraction, the proposed system shows significant improvement over the performance of the missing-data ASR using the speech separation mask. The improvement in recognition accuracy at SNRs $\leqslant 5$ dB are statistically significant (p-value < 0.05). For example, a maximum reduction in WER of 42% is obtained at the SNR of 5 dB. Note that, at SNRs $\geqslant 5$ dB, the performance of the proposed system is close to that of the missing-data recognizer with the ideal binary mask. Also note that both the proposed system and the missing-data ASR substantially outperform the conventional ASR with no preprocessing. Finally, comparing the results in Figs. 5 and 4, we can see that masks generated by the Hu and Wang system yield substantially better recognition results than those generated by spectral subtraction. This shows the better ability of the Hu and Wang system for speech segregation.

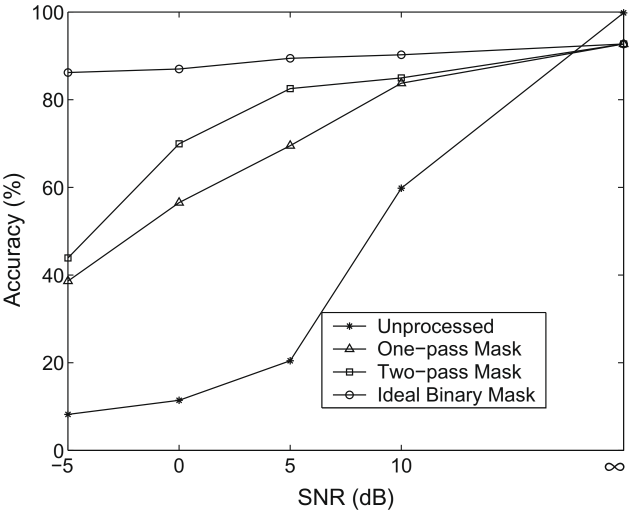

Fig. 5. Performance of the proposed system and the missing-data recognizer using the mask produced by the speech separation system of Hu and Wang (2004). Two-pass Mask and One-pass Mask refer to the performance of the proposed system and that of the missing-data ASR using the mask from Hu and Wang (2004), respectively. See Fig. 4 caption for other notations.

4. Concluding remarks

It is known that lattice re-scoring can be an order of magnitude faster than normal recognition (Young et al., 2000). Due to the small size of the lattice generated in our first pass, only a small increase in the computation time is observed for our system over that of the missing-data recognizer. We wish to emphasize that the computation time of the two-pass system is significantly lower than a purely top-down system such as the one suggested by Barker et al. (2005). The amount of savings in computation depends on the size of the lattice generated during the first pass of our system. While enlarging the lattice size ensures that the correct hypothesis is not pruned away during the first pass, it comes at the cost of increasing the number of states that examine a *T–F* unit labeled 2 in the second pass. We have observed that retaining 2 best hypotheses in each state provides a reasonable trade-off. With this choice of lattice pruning, our overall system is about five times faster than *the purely* top-down approach *of* Barker et al. (2005). Note that there is no performance degradation due to the use of the two-pass approach compared to the purely top-down approach, if adequate numbers of hypotheses are retained in the first pass.

In general, in a purely top-down approach, the search for alternate labels for each *T–F* unit would be exponential. While theoretically such an approach could achieve excellent recognition results, practically, the prohibitive computational cost would make it impossible. Barker et al. (2005) handle this problem by analyzing the output of the spectral subtraction mask in four frequency bands and employing an efficient token-passing search algorithm. However, the computational cost is still substantial because the search involves all states in the complete decoding lattice. On the other hand, the use of the bottom-up speech separation stage in our system enables us to create a small lattice containing only the most promising hypotheses. The allows the top-down hypothesis testing state to perform an efficient search for labeling a *T–F* unit.

The "evidence" measure used in the top-down hypothesis testing stage of our system is similar to state-level posterior probability proposed by Williams and Renals (1999). One could therefore use such local confidence measures also to create a hypothesis test. If the right threshold is chosen, we expect the results of such a hypothesis test to be similar to that provided by (6).

In our experiments, we have used masks produced by spectral subtraction and by a monaural CASA system as bottom-up speech separation masks. However, masks produced by other speech separation methods, including those produced through binaural processing (e.g. (Roman et al., 2003)), could also be used. Using a fixed threshold with spectral subtraction to generate candidate *T–F* units during the second pass, the system is unable to produce adequate *T–F* units labeled 2 at very low SNRs. This may explain the relatively low recognition accuracy obtained under these conditions. Additionally, fixing the number of tokens to 2 also limits the performance improvement. The optimal number of tokens may also depend on the vocabulary size. Future work will need to address these issues.

To conclude, the missing-data recognizer shows excellent recognition performance when the ideal binary mask is accurately estimated. Binary masks estimated by speech separation systems, however, contain a significant number of errors even under moderate amounts of additive noise. In this paper, we have presented a two-pass missing-data recognition system that refines the mask generated through front-end processing and provides significant reduction in the WER compared to that of the missing-data recognizer using the speech separation mask. Additionally, the proposed system outperforms a conventional ASR system with no preprocessing and prepocessing with spectral subtraction by a large margin. In our future work, we plan to also compare the performance with other noise reduction algorithms on the AURORA noisy speech recognition task (Pearce and Hirsch, 2000). It is worth noting that the system does not require a noise model. Hence, it is applicable under various noise conditions. Additionally, with the use of CASA systems for speech separation, bottom-up mask generation is independent of the recognition task (Srinivasan, 2006).

Acknowledgments

This research was supported in part by an AFOSR grant (FA9550-08-1-0155), an NSF grant (IIS-0534707). We thank J. Barker for help with the speech fragment decoder and E. Fosler-Lussier for helpful discussions. A preliminary version of this work was presented in 2005 ICASSP (Srinivasan and Wang, 2005a).

References

Barker, J.P., Cooke, M.P., Ellis, D.P.W., 2005. Decoding speech in the presence of other sources. Speech Communication 45, 5–25.

Bishop, C.M., 1994. Novelty detection and neural network validation. IEEE Proceedings of the Vision, Image and Signal processing 141 (4), 217–222.

Boersma, P., Weenink, D., 2002. Praat: doing phonetics by computer. Version 4.0.26. Last viewed on 24 October 2007. URL <http://www.fon.hum.uva.nl/praat>.

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

Bregman, A.S., 1990. Auditory Scene Analysis. The MIT Press, Cambridge, MA.

Brown, G.J., Barker, J., Wang, D.L., 2001. A neural oscillator sound separator for missing data speech recognition. In: Proceedings of the International Joint Conference on Neural Networks '01, pp. 2907–2912.

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

Darwin, C.J., 2008. Listening to speech in the presence of other sounds. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 1011–1021.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP 28 (4), 357–366.

Droppo, J., Acero, A., Deng, L., 2002. A nonlinear observation model for removing noise from corrupted speech log mel-spectral energies. In: Proceedings of the International Conference on Spoken Language Processing '02, pp. 1569–1572.

Drygajlo, A., El-Maliki, M., 1998. Speaker verification in noisy environment with combined spectral subtraction and missing data theory. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing '98, vol. 1, pp. 121–124.

Ephraim, Y., 1992. A Bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Transactions on Signal Processing 40 (4), 725–735.

Gales, M., Young, S., 2007. The application of Hidden Markov models in speech recognition. Foundations and Trends in Signal Processing 1 (3), 195–304.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Communication 16, 261–291.

Hu, G., Wang, D.L., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Transactions on Neural Networks 15, 1135–1150.

Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing. Prentice-Hall, PTR, Upper Saddle River, NJ.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing '84, pp. 111–114.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. Taylor and Francis, Boca Raton, FL.

Markou, M., Singh, S., 2003. Novelty detection: a review–part 1: statistical approaches. Signal Processing 83 (12), 2481–2497.

McCowan, I., Bourlard, H., 2003. Microphone array post-filter based on noise field coherence. IEEE Transactions on Speech and Audio Processing 11 (6), 709–716.

McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, NY, NY.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1988. An efficient auditory filterbank based on the gammatone function. Applied Psychology Unit (APU) Report 2341.

Pearce, D., Hirsch, H., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proceedings of the Sixth International Conference on Spoken Language Processing '00, pp. 29–32.

Renevey, P., Drygajlo, A., 2001. Detection of reliable features for speech recognition in noisy conditions using a statistical criterion. In: Proceedings of the Consistent and Reliable Acoustic Cues for Sound Analysis Workshop '01, pp. 71–74.

Roman, N., Wang, D.L., Brown, G.J., 2003. Speech segregation based on sound localization. The Journal of the Acoustical Society of America 114, 2236–2252.

Seltzer, M.L., Raj, B., Stern, R.M., 2000. Classifier-based mask estimation for missing feature methods of robust speech recognition. In: Proceedings of the International Conference on Spoken Language Processing '00, pp. 538–541.

Srinivasan, S., 2006. Integrating computational auditory scene analysis and automatic speech recognition. Ph.D. Thesis, Biomedical Engineering Department, The Ohio State University.

Srinivasan, S., Roman, N., Wang, D.L., 2006. Binary and ratio time–frequency masks for robust speech recognition. Speech Communication 48, 1486–1501.

Srinivasan, S., Wang, D.L., 2005a. Robust speech recognition by integrating speech separation and hypothesis testing. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing '05, vol. 1, pp. 89–92.

Srinivasan, S., Wang, D.L., 2005b. A schema-based model for phonemic restoration. Speech Communication 45, 63–87.

Stark, H., Woods, J.W., 2002. Probability and Random Processes with Applications to Signal Processing, third ed. Prentice-Hall, Upper Saddle River, NJ.

Tax, D.M.J., Duin, R.P.W., 1998. Outlier detection using classifier instability. In: Amin, A., Dori, D., Pudil, P., Freeman, H. (Eds.), Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition, Lecture Notes in Computer Science, vol. 1451, Springer, Berlin, pp. 593–601.

van Hamme, H., 2004. Robust speech recognition using cepstral domain missing data techniques and noisy masks. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing '04, vol. 1, pp. 213–216.

Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing '90, pp. 845–848.

Varga, A.P., Steeneken, H.J.M., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK.

Wang, D.L., Brown, G.J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Transactions on Neural Networks 10 (3), 684–697.

Wang, D.L., Brown, G.J. (Eds.), 2006. Computational Auditory Scene Analysis: Principles, Algorithms and Applications. Wiley-IEEE Press, Hoboken, NJ.

Williams, G., Renals, S., 1999. Confidence measures from local posterior probability estimates. Computer Speech and Language 13, 395–413.

Young, S., Kershaw, D., Odell, J., Valtchev, V., Woodland, P., 2000. The HTK Book (for HTK Version 3.0). Microsoft Corporation.