INCORPORATING THE VOICING INFORMATION INTO HMM-BASED AUTOMATIC SPEECH RECOGNITION

Peter Jančovič and Münevver Köküer

Electronic, Electrical & Computer Engineering, University of Birmingham, Birmingham, UK

{p.jancovic, m.kokuer}@bham.ac.uk

ABSTRACT

In this paper, we propose a novel model for incorporating the voicing information in a speech recognition system. The voicing information employed is estimated by a novel method that can provide this information for each filter-bank channel, without requiring any information about the fundamental frequency. A Viterbi-style training procedure is employed to estimate the voicing-probability of each mixture at each HMM state. Experiments are performed on noisy speech data from the Aurora 2 database. Significant performance improvements are achieved at low SNRs when the voicing information is incorporated within the standard model and two models that had already compensated for the effect of the noise.

Index Terms— Source-filter model, voicing estimation, speech recognition, HMM, acoustic modeling, voicing probability, noise robustness, missing-feature model, multi-conditional training, Aurora 2 database

1. INTRODUCTION

Much effort has been devoted to finding an effective representation of speech signals for automatic speech recognition. Current frame-based speech representations, with the mel-frequency cepstral coefficients (MFCCs) [1] and frequency-filtered logarithm filter-bank energies [2] being among the most successful, typically aim at describing the envelope of a short-time spectra, which corresponds to the characteristic of the vocal-tract filter. However, speech sounds are produced by passing a source-signal through a vocal-tract filter, i.e., different speech sounds may be produced when a given vocal-tract filter is excited by different source-signals. Thus, a more appropriate representation and modeling of speech signals should include both the information about the vocal-tract filter and the source-signal. The information about the source-signal may be characterized by a voicing character of a speech frame or individual frequency bands and the value of the fundamental frequency (F0). Our study in this paper is concerned with the incorporation of the voicing information.

There has been several works investigating the incorporation of the source-signal information into speech recognition. The authors in [3] [4] [5] [6] [7] investigated the use of various measures for estimating the voicing-level of a speech frame and appended these voicing features into the feature representation. In addition to voicing features, the information on F0 was employed in [4] [5]. In [3] the effect of including the voicing features under various training procedures was also studied. Experiments in the above papers were performed only on speech signal not corrupted by an additional noise and modest improvements have been reported. In [8], the voicing information was included by decomposing speech signal into simultaneous periodic and aperiodic streams and weighting the contribution of each stream during the recognition. This method requires information about the fundamental frequency. Significant improvements on noisy speech recognition on Aurora 2 connected-digit database have been demonstrated, however, these results were achieved by using the F0 estimated from the clean speech. The authors in [9] divided phoneme-based models of speech into a subset of voiced and unvoiced models and used this division to restrict the Viterbi search during the recognition. The effect of such division of models itself was not presented. In [10] an HMM model was estimated based only on high-energy frames, which effectively corresponds to the voiced speech. This was observed to improve the performance in noisy conditions.

In this paper, we propose a novel model for incorporating the voicing information in an automatic speech recognition (ASR) system. Our model differs from the above works in the following: i) the voicing information employed is estimated by a novel method that can provide this information for each filter-bank channel, while requiring no information about the F0; ii) the voicing-information is incorporated within an HMM-based statistical framework in the back-end of the ASR system; iii) the evaluation is performed on noisy speech recognition. Note that the method for estimation of the voicing information was introduced in [11] and further analysis and evaluations of the method were presented in [12]. While in [11] [12] the voicing information was employed as a mask in a missing-feature ASR system that modeled only the characteristics of the vocal-tract filter, in this paper we present modeling of the voicing information within an HMM-based ASR system. In the proposed model, having the trained HMMs, each mixture at each HMM state is associated with a voicing-probability, which is estimated by a separate Viterbi-style training procedure (without altering the trained HMMs). The incorporation of the voicing-probability serves as a penalty during recognition for those mixtures/states whose voicing information does not correspond to the voicing information of the signal. The effect of employing the voicing-probability about an entire frame and about each filter-bank channel is also explored. The incorporation of the voicing information is evaluated in a standard model and in two models that had compensated for the effect of the noise, missing-feature (e.g., [13]) and multi-conditional training. Experiments are performed on the Aurora 2 database. Experimental results show significant improvements in recognition performance in strong noisy conditions achieved by the models incorporating the voicing information.

2. ESTIMATING THE VOICING INFORMATION OF FILTER-BANK CHANNELS

The estimation of the voicing information of speech signal for each filter-bank channel is performed by algorithm we introduced in

[11] and presented with further analyses in [12]. It exploits the quasi-periodicity of voiced speech signals and the effect of short-time processing – due to these, the shape of short-time magnitude spectra of voiced speech around each harmonic frequency should follow approximately the shape of the magnitude spectra of the frame analysis window. Note that it does not require any information about the fundamental frequency. It has been demonstrated that the voicing information of filter-bank channels can be detected with 5% false-acceptance and false-rejection accuracy at 10dB local SNR [12]. Below are the steps of the method:

1) Short-time magnitude-spectra calculation: A frame of a time-domain signal is weighted by a frame-analysis window function, expanded by zeros and the FFT is applied to provide a short-time magnitude-spectra.

2) Voicing-distance calculation: For each peak of the signal short-time magnitude-spectra, a distance, referred to as *voicing-distance* $vd(k)$, between the spectra around the peak and magnitude-spectra of the frame window is computed, i.e.,

$$vd(k_p) = \left[\frac{1}{2M+1}\sum_{m=-M}^{M}\Big(|S(k_p+m)| - |W(m)|\Big)^2\right]^{1/2} \quad (1)$$

where k_p is frequency-index of a spectral peak and M determines the number of components of the spectra at each side around the peak to be compared. The spectra of the signal, $S(k)$, and frame-window, $W(k)$, are normalized to have magnitude value equal to 1 at the peak prior to their use in Eq. 1.

3) Voicing-distance calculation for filter-bank channels: The voicing-distance for each filter-bank channel is calculated as a weighted average of the voicing-distances within the channel, reflecting the calculation of filter-bank energies that are used to derive features for recognition, i.e.,

$$vd^{fb}(b) = \frac{1}{X(b)} \cdot \sum_{k=k_b}^{k_b+N_b-1} vd(k) \cdot G_b(k) \cdot |S(k)|^2 \quad (2)$$

where $G_b(k)$ is the frequency-response of the filter-bank channel b, and k_b and N_b are the lowest frequency-component and number of components of the frequency response, respectively. The $X(b) = \sum_{k=k_b}^{k_b+N_b-1} G_b(k)|S(k)|^2$, i.e., the overall filter-bank energy value.

4) Postprocessing of the voicing-distances: The voicing-distance obtained from Eq. 1 and Eq. 2 were filtered by 2D median filters in order to eliminate accidental errors.

The voicing information of a filter-bank channel could be directly expressed by the voicing-distance value. However, for simplicity of its incorporation, in this paper, a binary valued voicing information was used. A filter-bank channel b is considered as voiced, i.e., $v(b) = 1$, if the corresponding voicing-distance $vd^{fb}(b)$ is below a given threshold (based on [12] the value 0.21 was used) and unvoiced, i.e., $v(b) = 0$, otherwise. Note that in experimental evaluation presented in Section 4, we also used the voicing-information about an entire frame; a frame is assigned as voiced if there are at least three filter-bank channels detected as voiced.

3. INCORPORATING THE VOICING INFORMATION INTO AN HMM-BASED ASR SYSTEM

This section presents the proposed incorporation of the voicing information, estimated in Section 2, in the back-end of speech recognition system. The voicing-probability is estimated by a separate Viterbi-style training procedure that is performed after the HMMs have been trained (i.e., the trained HMMs are not altered). The following sections give detailed description of the proposed method and discuss the effect of incorporation of the voicing-probability during the state-time recognition search.

3.1. Incorporating the voicing information during recognition

During the recognition, the standard HMM state emission probability of a spectral feature-vector $\mathbf{y}_t$ at frame-time t in state s, i.e., $P(\mathbf{y}_t|s)$, is replaced by calculating the joint probability of the spectral feature vector and the voicing vector $\mathbf{v}_t$, i.e., $P(\mathbf{y}_t, \mathbf{v}_t|s)$. Considering that all spectral features and voicing features are independent of one another, using L mixture densities the $P(\mathbf{y}_t, \mathbf{v}_t|s)$ is calculated in the proposed model as

$$P(\mathbf{y}_t, \mathbf{v}_t|s) = \sum_{l=1}^{L} P(l|s) \prod_b P(y_t(b)|l,s)P(v_t(b)|l,s) \quad (3)$$

where $P(l|s)$ is the weight of the l^{th} mixture component, and $P(y_t(b)|l,s)$ and $P(v_t(b)|l,s)$ are the probability of the b^{th} spectral feature and voicing feature, respectively, given state s and mixture l. Note that instead of using the voicing information of each filter-bank channel as considered above, one may use only information about frame voicing. Experiments were performed with both levels of the voicing information.

3.2. Estimating the voicing-probability for HMM states

The estimation of the voicing-probability $P(\mathbf{v}|l,s)$ at each HMM state and mixture was performed by a Viterbi-style training procedure using the training data-set.

Given a speech utterance, for each frame t we have the spectral-feature vector $\mathbf{y}_t$ and voicing vector $\mathbf{v}_t$, resulting a sequence of $\{(\mathbf{y}_1, \mathbf{v}_1), \ldots, (\mathbf{y}_T, \mathbf{v}_T)\}$. The Viterbi algorithm is then used to obtain the state-time alignment of the sequence of feature vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ on the HMMs corresponding to the speech utterance. This provides an association of each feature vector $\mathbf{y}_t$ to some HMM state s. The posterior probability that the mixture-component l (at the state s) have generated the feature vector $\mathbf{y}_t$ is then calculated as

$$P(l|\mathbf{y}_t, s) = \frac{P(\mathbf{y}_t|l,s)P(l|s)}{\sum_{l'} P(\mathbf{y}_t|l',s)P(l'|s)} \quad (4)$$

where the mixture-weight $P(l|s)$ and the probability density function of the spectral features used to calculate the $P(\mathbf{y}_t|l,s)$, are obtained as an outcome of the HMM training.

For each mixture l and HMM state s, we collect (over the entire training data-set) the posterior probabilities $P(l|\mathbf{y}_t, s)$ for all $\mathbf{y}_t$'s associated with the state s together with the corresponding voicing vectors $\mathbf{v}_t$'s. The voicing-probability of the b^{th} feature can then be obtained as

$$P(v(b) = a|l,s) = \frac{\sum_{t:\mathbf{y}_t \in s} P(l|\mathbf{y}_t, s) \cdot \delta(v_t(b), a)}{\sum_{t:\mathbf{y}_t \in s} P(l|\mathbf{y}_t, s)} \quad (5)$$

where $a \in \{0, 1\}$ is the value of voicing and $\delta(v_t(b), a)$=1 when $v_t(b)$=a, otherwise zero.

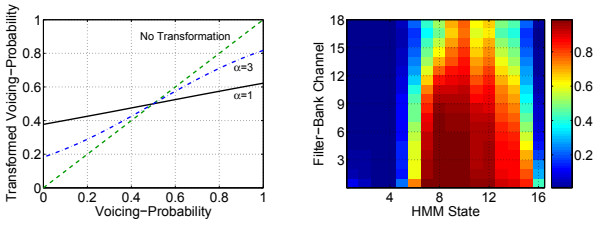
3.3. Transformation of the voicing-probability

In the overall probability calculation by Eq. 3, the value of the voicing-probability may need to be scaled, since it is real probability

while the first term in the product in Eq. 3 is a likelihood (i.e., they have a different range of values). This can be performed by employing a sigmoid function to transform the $P(v(b)|l,s)$ for each b to a new value, i.e.,

$$P(v(b)|l,s) = \frac{1}{1+e^{-\alpha(P(v(b)|l,s)-0.5)}} \quad (6)$$

where α is a constant defining the slope of the function and the value 0.5 gives shift of the function. An example of voicing-probability transformation with various values for α and no transformation case are depicted on Figure 1(a). The bigger the value of α is the greater the effect of the voicing-probability on the overall probability. An appropriate value for α can be decided based on a small set of experiments on a development data. In our experiments, values of α within the range $<1,3>$ yielded the best recognition results. The α is set to 1.5 for feature-level (6 for frame-level) voicing for all the experiments presented in the paper.



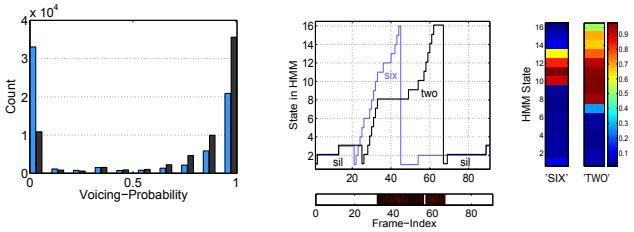
(a) (b)

Fig. 1. *Voicing-probability transformation by using a sigmoid function with various slope parameter α and by using no transformation (a). An example of the estimated voicing-probability for a 16 state HMM model of word 'five' (b).*

An example of the estimated voicing-probability for an HMM model of word 'five' is depicted on Figure 1(b). It can be seen that, for instance, the first four states have a low probability of being voiced over the entire frequency range, which may correspond to the unvoiced phoneme $\backslash f \backslash$.

3.4. The effect of the voicing-probability during the recognition

This section demonstrates the effect of incorporating the voicing-probability on the recognition process. A frame-level voicing information was considered for simplicity of presentation of the results. An experiment was performed to identify the amount of disagreement between the voicing information of models and the signal. For each voiced frame of the signal, the voicing-probability of the state the frame is associated to according to the best path through the state-time trellis found by the Viterbi algorithm is obtained. The histograms of these voicing-probabilities collected over noisy test speech utterances (white noise at 0dB) are depicted on Figure 2(a). It can be seen that when the voicing information is not incorporated (blue) there is a large amount of voiced frames being assigned to states with low voicing-probability. This situation is significantly improved when the voicing information is incorporated since this acts as a penalty during the recognition for those states whose voicing is not in agreement with the voicing of the signal. Figure 2(b) shows an example of the Viterbi-found path for a speech utterance "two" without and with using the voicing-probability, resulting in being recognized as "six" and "two", respectively, together with the estimated voicing information for each frame of the utterance. A significant disagreement between the voicing of the model and signal can be seen when the voicing is not incorporated, e.g., voiced frames after frame-index 43 are assigned to the silence model.



(a) (b)

Fig. 2. *Histogram of state voicing-probabilities associated with voiced frames without (blue) and with (black) using the voicing information (a). Recognition of a speech utterance "two", and state-time path, without (blue) and with (black) using voicing-probability. Below: frame-level voicing of the utterance. Right: voicing-probability of each state for HMM of digit "six" and "two" (b).*

4. EXPERIMENTAL RESULTS

The experiments were carried out on the Aurora 2.0 English language connected-digit database [14]. The frequency-filtered (FF) logarithm filter-bank energies [2] were used as speech feature representation, due to their suitability for missing-feature based recognition. Note that the FF-features achieved similar performance (in average) as standard MFCCs. The FF-features were obtained with the following parameter set-up: frames of 32 ms length with an overlap of 10 ms between frames were used; both preemphasis and Hamming window were applied to each frame; the short-time magnitude spectra, obtained by applying the FFT, was passed to Mel-spaced filter-bank analysis with 20 channels; the obtained logarithm filter-bank energies were filtered by using the filter $H(z)=z-z^{-1}$ [2]. A feature vector consisting of 18 elements was obtained (the edge values were excluded). An FF-feature was assigned as voiced (i.e., $v(b)$=1) only if both of the filter-bank channels involved in the calculation of the FF-feature were voiced, and unvoiced otherwise. In order to include dynamic spectral information, the first-order delta parameters were added to the static FF-feature vector. A continuous-observation left-to-right HMM with 16 states (no skip allowed) was used to model each digit; the pdf at each state was modeled with three and ten Gaussian mixtures when using clean training and multi-conditional training, respectively, and diagonal covariance matrices. The training of HMMs was performed on utterances from the training set. The noisy speech data from the Set A in Aurora 2.0 were used for recognition experiments. The results for clean speech were omitted from all figures as marginal differences were observed by incorporating the voicing-probability (similar observations were reported also in [5]).

The evaluation of the proposed model for voicing incorporation is first performed using a standard model trained on clean data. Results, presented in Figure 3, show that incorporation of the voicing-probability provides significant improvement of the recognition accuracy at low SNRs in all noisy conditions. It was observed that the voicing-probability incorporation caused an increase of insertions in the case of Babble noise, which is due to this noise being a background speech. This could be improved by modifying the word insertion penalty or employing a speech-of-interest detection. The

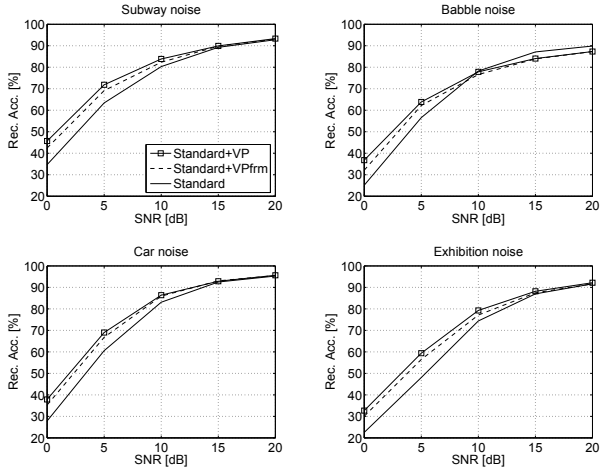


Fig. 3. *Recognition accuracy results obtained by the standard model without and with incorporated voicing probability.*

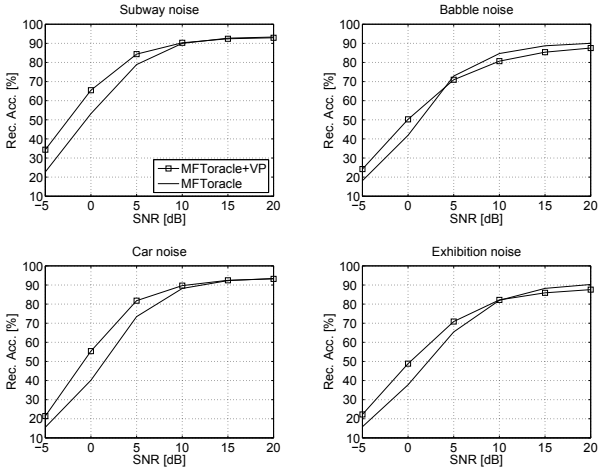


Fig. 4. *Recognition accuracy results obtained by the MFT-model using the oracle mask without and with incorporating the voicing probability.*

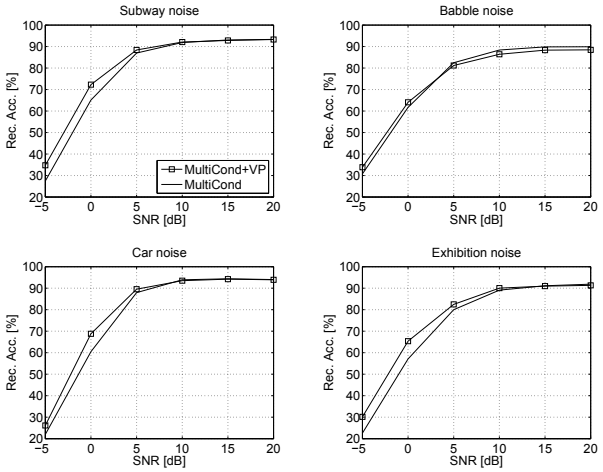


Fig. 5. *Recognition accuracy results obtained by the multi-conditional trained model without and with incorporating the voicing probability.*

former was used here – this resulted in slight decrease of the performance at high SNRs, however, provided significant improvements at low SNRs. Figure 3 shows that the incorporation of the voicing-probability on the feature-level gives in all noisy conditions slightly better results than using the frame-level voicing (standard+VPfrm), which is a consequence of a more detailed voicing information modeling. Note that the use of a frame-level voicing may be more deficient (against feature-level) in a more difficult task, i.e., larger vocabulary system, which is currently under our investigation.

Next, evaluations were performed on two types of models that had compensated for the effect of noise – this was conducted in order to determine whether the incorporation of the voicing information can still provide improvements (as employment of a noise compensation would effectively decrease the amount of misalignment of voicing). The first noise-compensated model was based on the missing-feature theory (MFT). We used the marginalization-based MFT model. In order to obtain the best (idealized) noise compensation, this model employs the oracle mask, obtained based on the full a-priori knowledge of the noise. Specifically, the static features whose local SNR is below 0dB were marginalized. Experimental results are presented in Figure 4. The second type of noise-compensated model was obtained by using the multi-conditional training. Experimental results are presented in Figure 5. It can be seen from Figure 4 and Figure 5 that the incorporation of the voicing-probability did not improve the performance at high SNRs, which may be due to the effectiveness of the noise-compensation. The decrease at high SNRs in the case of Babble noise (and Exhibition noise in Figure 4) is, similarly as in the standard model discussed earlier, due to the increased insertions. However, it can be seen that at low SNRs, even the noise effect had already been compensated, the incorporation of the voicing-probability within each type of the noise-compensated models provides significant improvements in the recognition accuracy.

5. CONCLUSION

In this paper, we presented a novel model for speech recognition that incorporates the voicing information of speech signal. A Viterbi-style training procedure for estimation of the voicing-probability for each mixture at each HMM state was presented. The effectiveness of the method was demonstrated within a standard model and two types of noise-compensated models, missing-feature and multi-conditional training. Experimental evaluation was performed on noisy speech data from the Aurora 2 database. Significant performance improvements were observed at strong noisy conditions when the voicing information is incorporated in the standard model, and also in both models which had already compensated for the effect of noise.

ACKNOWLEDGEMENT

This work was supported by UK EPSRC grant EP/D033659/1.

6. REFERENCES

[1] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustic, Speech, and Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

[2] C. Nadeu, D. Macho, and J. Hernando, "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, vol. 34, pp. 93–114, 2001.

[3] D.L. Thomson and R. Chengalvarayan, "The use of voicing features in HMM-based speech recognition," *Speech Communication*, vol. 37, pp. 197–211, 2002.

[4] A. Ljolje, "Speech Recognition Using Fundamental Frequency and Voicing in Acoustic Modeling," *ICSLP, Denver, USA*, pp. 2137–2140, 2002.

[5] N. Kitaoka, D. Yamada, and S. Nakagawa, "Speaker Independent Speech Recognition using Features based on Glottal Sound Source," *ICSLP, Denver, USA*, pp. 2125–2128, 2002.

[6] A. Zolnay, R. Schluter, and H. Ney, "Extraction methods of voicing feature for robust speech recognition," *Eurospeech, Geneva, Switzerland*, pp. 497–500, 2003.

[7] M. Graciarena, H. Franco, J. Zheng, D. Vergyri, and A. Stolcke, "Voicing Feature Integration in SRI's Decipher LVCSR System," *ICASSP, Montreal, Canada*, vol. I, pp. 921–924, 2004.

[8] P.J.B. Jackson, D.M. Moreno, M.J. Russell, and J. Hernando, "Covariation and weighting of harmonically decomposed streams for ASR," *Eurospeech, Geneva, Switzerland*, pp. 2321–2324, 2003.

[9] D. O'Shaughnessy and H. Tolba, "Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision," *ICASSP, Phoenix, Arizona*, vol. I, pp. 413–416, 1999.

[10] P. Jančovič and J. Ming, "Combining the union model and missing feature method to improve noise robustness in ASR," *ICASSP, Orlando, Florida*, vol. I, pp. 69–72, 2002.

[11] P. Jančovič and M. Köküer, "Voicing-Character Estimation of Speech Spectra: Application to Noise-Robust Speech Recognition," *ICASSP, Toulouse, France*, pp. 257–260, 2006.

[12] P. Jančovič and M. Köküer, "Estimation of Voicing-Character of Speech Spectra based on Spectral Shape," *IEEE Signal Processing Letters*, vol. 14, no. 1, pp. 66–69, Jan. 2007.

[13] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[14] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA ITRW ASR'2000: Challenges for the New Millenium, Paris, France*, Sept. 2000.