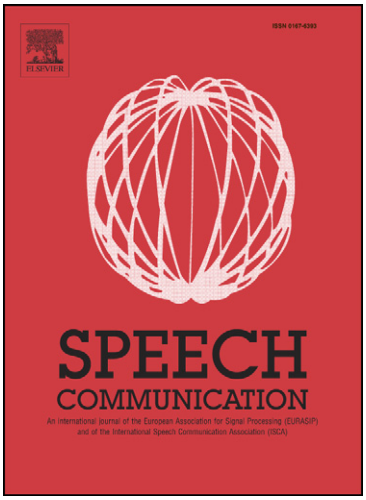

ISSN 0167-6393
ELSEVIER
SPEECH
COMMUNICATION
An International Journal of the European Association for Signal Processing (EURASIP)
and of the International Speech Communication Association (ISCA)

Incorporating the Voicing Information into HMM-based Automatic Speech Recognition in Noisy Environments

Peter Jančovič, Münevver Köküer

PII: S0167-6393(09)00005-3
DOI: 10.1016/j.specom.2009.01.003
Reference: SPECOM 1774

To appear in: *Speech Communication*

Received Date: 29 May 2008
Revised Date: 10 December 2008
Accepted Date: 20 January 2009

Please cite this article as: Jančovič, P., Köküer, M., Incorporating the Voicing Information into HMM-based Automatic Speech Recognition in Noisy Environments, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.01.003

Incorporating the Voicing Information into HMM-based Automatic Speech Recognition in Noisy Environments

Peter Jančovič *, Münevver Köküer

Electronic, Electrical & Computer Engineering, University of Birmingham,
Pritchatts Road, B15 2TT Birmingham, UK
Tel: +44 121 4144316; Fax: +44 121 4144291

Abstract

In this paper, we propose a model for the incorporation of voicing information into a speech recognition system in noisy environments. The employed voicing information is estimated by a novel method that can provide this information for each filter-bank channel and does not require information about the fundamental frequency. The voicing information is modelled by employing the Bernoulli distribution. The voicing model is obtained for each HMM state and mixture by a Viterbi-style training procedure. The proposed voicing incorporation is evaluated both within a standard model and two other models that had compensated for the noise effect, the missing-feature and the multi-conditional training model. Experiments are first performed on noisy speech data from the Aurora 2 database. Significant performance improvements are achieved when the voicing information is incorporated within the standard model as well as the noise-compensated models. The employment of voicing information is also demonstrated on a phoneme recognition task on the noise-corrupted TIMIT database and considerable improvements are observed.

Key words: Source-filter model, voicing estimation, HMM, automatic speech recognition, voicing modelling, noise robustness, missing feature, Aurora 2, phoneme recognition

1 Introduction

Speech sounds are produced by passing a source-signal through a vocal-tract filter (Fant, 1960), i.e., different speech sounds may be produced when a given vocal-tract filter is excited by different source-signals. Thus, the representation and modelling of speech signals should include information about both the vocal-tract filter and the source-signal. The characteristics of the vocal-tract filter are reflected by the envelope of a short-time spectrum. As the source-signal may in general consist of a mixture of white noise and train pulses with a period corresponding to the fundamental frequency (F0), the information about the source-signal may be characterised by the voicing character (i.e., voiced/unvoiced) of individual frequency-regions and the value of the fundamental frequency.

Current frame-based speech representations for speech pattern processing – with the mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) and the frequency-filtered logarithm filter-bank energies (Nadeu et al., 2001) being among the most successful – typically aim at representing the characteristics of the vocal-tract filter. The use of voicing information in

* Corresponding author.
Email addresses: p.jancovic@bham.ac.uk (Peter Jančovič),
m.kokuer@bham.ac.uk (Münevver Köküer).

speech recognition was suggested in the 1970s (Rabiner and Sambur, 1976). However, until recently this has mainly been for speech end-point detection. The employment of the fundamental frequency in speech recognition may be of interest especially for tonal languages and this has recently been investigated in, e.g., Huang and Seide (2000). The incorporation of the voicing-information in a speech recognition system is of concern in this paper.

Recently there have been several works investigating the incorporation of the voicing information into speech recognition. The authors in (Thomson and Chengalvarayan, 2002) (Ljolje, 2002) (Kitaoka et al., 2002) (Zolnay et al., 2003) (Graciarena et al., 2004) have investigated the use of various measures for estimating the level of voicing of an entire speech frame and have appended these voicing features into a standard spectral-envelope feature representation. The voicing features employed were obtained based on an autocorrelation-function (Thomson and Chengalvarayan, 2002) (Zolnay et al., 2003) (Graciarena et al., 2004), energy of the residual error signal from the linear prediction analysis (Kitaoka et al., 2002), a harmonic product spectrum (Zolnay et al., 2003), and an entropy of high-order cepstrum (Graciarena et al., 2004). In addition to the voicing features, the information on F0 was employed in both Ljolje (2002) and Kitaoka et al. (2002). In Thomson and Chengalvarayan (2002), the effect of including the voicing features under various training procedures was also studied. Experimental evaluations in the above mentioned works were presented only on a speech signal that was not corrupted by additional noise and mainly modest improvements have been reported. Beaufays et al. (2003) built a soft speech/non-speech detector by employing several features, including a frame-level voicing, and used the output of the detector to penalise speech/non-speech confusions between the models and the signal.

In Jackson et al. (2003), the voicing information was included by decomposing the speech signal into simultaneous periodic and aperiodic streams and weighting the contribution of each stream during the recognition. This method requires knowledge about the fundamental frequency. Significant improvements on noisy speech recognition have been demonstrated on the Aurora 2 connected-digit database, however, these results were achieved by using the F0 estimated from clean speech. Similar decomposition approach but employing comb filters independently designed in each sub-band was presented in Ishizuka et al. (2006) with recognition accuracy improvements reported for the Aurora 2J database when using the standard model trained on clean speech. In Jančovič and Ming (2002) an HMM model was estimated based only on high-energy frames, which effectively correspond to the voiced speech. This was observed to improve the performance on a digit recognition task in noisy conditions. O'Shaughnessy and Tolba (1999) divided the 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. Niyogi and Ramesh (2003) employed the voicing onset time in a two-pass HMM-based speech recognition system to reclassify the segments recognised as stop consonants.

In this paper we present a novel model for the incorporation of the voicing information into an HMM-based automatic speech recognition (ASR) system in noisy conditions. This paper extends our preliminary work presented in Jančovič and Köküer (2007b). 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. The voicing information is incorporated within an HMM-based statistical framework in the back-end of the ASR system. The Bernoulli distribution is employed to model the voicing information. The parameters of the voicing model associated with each mixture at each HMM state are 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. To deal with the effect of noise on voicing, marginalisation of the voicing information of unvoiced features during recognition is proposed. The effect of employing the voicing information for an entire frame and for each filter-bank channel is demonstrated. Appending the voicing information into the feature vector is also undertaken and compared to the proposed model. The incorporation of the voicing information is evaluated in a standard model trained on clean speech and in models that compensate for the effect of the noise, including the missing-feature model (e.g., Cooke et al., 2001), and the multi-conditional training model. Experiments are performed in various noisy conditions and SNRs on connected-digit recognition on the Aurora 2 database and on the phoneme recognition on the TIMIT database. Experimental results show significant improvements in recognition performance in noisy conditions achieved by models with incorporated voicing information.

The paper is set out as follows: The method for estimation of the voicing information is presented in Section 2. The proposed incorporation of the voicing information into an HMM-based ASR system is described in Section 3. Experimental evaluations on connected-digit recognition and phoneme recognition are presented in Section 4, and conclusions are presented in Section 5.

2 Estimating voicing information of filter-bank channels

Estimation of voicing information of a speech signal for each filter-bank channel is performed by an algorithm we introduced previously in (Jančovič and Köküer, 2007a), where various analyses are also presented. This algorithm exploits the effect of short-time processing, due to which the shape of the short-time magnitude spectrum of voiced speech around each harmonic frequency should follow approximately the shape of the magnitude spectrum of the frame-analysis window. Note that it does not require any information about the fundamental frequency. The steps of the method are as follows:

1) Short-time magnitude-spectrum 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 signal short-time magnitude-spectrum, denoted by $S(k)$. The frame length of 256 samples (corresponding to 32 ms) and FFT of 512 samples was used here.

2) Voicing-distance calculation for spectral peaks: For each peak of the signal short-time magnitude-spectrum, a distance, referred to as *voicing-distance* $vd(k)$, between the spectrum around the peak and magnitude-spectrum of the frame-analysis window $W(k)$ is calculated as

$$vd(k_p) = \left[\frac{1}{2M+1} \sum_{m=-M}^{M} \left(\frac{|S(k_p+m)|}{|S(k_p)|} - \frac{|W(m)|}{|W(0)|} \right)^2 \right]^{1/2} \qquad (1)$$

where k_p is the frequency-index of a spectral peak and M determines the number of components of the spectrum at each side around the peak to be compared. The Hamming window was used and the parameter M was set to 2. The voicing-distance values for the frequency points other than peaks were obtained by an interpolation between the voicing-distance values of adjacent peaks, as described in (Jančovič and Köküer, 2007a).

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.,:

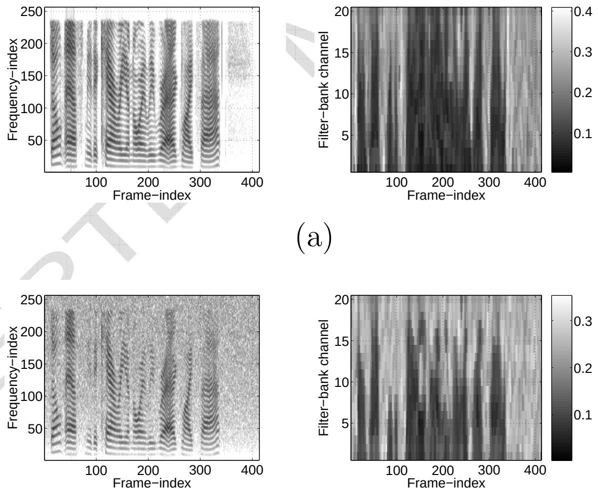
$$vd^{fb}(b) = \frac{1}{X(b)} \cdot \sum_{k=k_b}^{k_b+N_b-1} vd(k) \cdot G_b(k) \cdot |S(k)|^2 \tag{2}$$

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. The filters $G_b(k)$ here correspond to the Mel-spaced filter-bank analysis, which was also used in the feature extraction (see Section 4.1.2).

4) Postprocessing of the voicing-distances: The voicing-distances obtained from Eq. 1 (after the interpolation) and Eq. 2 were filtered by 2D median filters in order to eliminate accidental errors. Median filters of size 5x9 and 3x3 were used, respectively, where the first number corresponds to the number of frames.

The voicing information of a filter-bank channel could be directly expressed by the voicing-distance value. However, in this paper, for the simplicity of its incorporation, 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 otherwise it is considered unvoiced, i.e., $v(b) = 0$. The value of the threshold provides a trade-off between the amount of false-acceptance and false-rejection errors. The threshold range from 0.18 to 0.21 was shown to provide false-acceptance below 5% (Jančovič and Köküer, 2007a). This range provided similar recognition accuracy performance in experiments here and the presented results are obtained with the threshold value of 0.21. It should be noted that in the experimental evaluation presented in Section 4, we also used the voicing-information about an entire frame, where a frame is assigned as voiced if there are at least three filter-bank channels detected as voiced.

Figure 1 depicts examples of spectrograms of clean and noisy speech and the corresponding voicing distances for filter-bank channels. It can be seen that the harmonic regions obtain low voicing distance values.



(a)

(b)

Fig. 1. *Spectrogram (left) and corresponding voicing distances of filter-bank channels (right) of speech utterance which is clean (a) and corrupted by White noise at 15dB (b).*

3 Incorporating the voicing information into an HMM-based ASR system

This section presents the incorporation of the voicing information, estimated in Section 2, within an HMM-based speech recognition system. The aim of the incorporation of the voicing information is to penalise during the recognition those HMM states whose voicing model is in disagreement with the voicing information of the signal being recognised. The following sections give detailed descriptions of the estimation of voicing models, the incorporation of the voicing information and control of its effect during recognition, and the demonstration of the effect of the incorporated voicing information during the state-time recognition search.

3.1 Estimating the voicing models for HMM states

Let $\mathbf{v} = (v(1), \ldots, v(B))$ denote the voicing information vector at a given frame, where $v(b)$ is the voicing information of the channel b and B is the number of channels. The voicing-probability $P(\mathbf{v}|l, s)$ for each HMM state s and mixture l is modelled using a multivariate Bernoulli distribution as

$$P(\mathbf{v}|l, s) = \prod_{b=1}^{B} \mu_{b,l,s}^{v(b)} (1 - \mu_{b,l,s})^{1-v(b)}. \tag{3}$$

The parameter $\mu_{b,l,s}$ of the distribution can be estimated using a Baum-Welch or Viterbi training procedure. The latter was used in this paper. On the training data-set, a separate Viterbi-style training procedure was performed after the HMMs have been trained using spectral features, i.e., the trained HMMs are not altered. The following gives details of the voicing model estimation.

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 state s) have generated the feature vector $\mathbf{y}_t$ is then calculated as

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

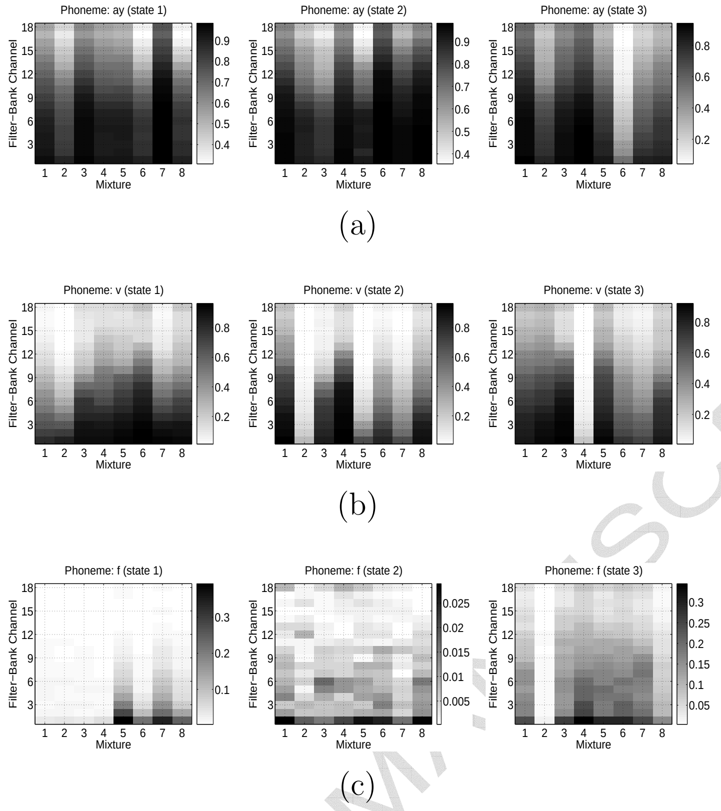
where the mixture-weight $P(l|s)$ and the probability density function of the spectral features used to calculate the $P(\mathbf{y}_t|l, s)$, are obtained as an outcome of the HMM training.

For each mixture l and HMM state s, we collect (over the entire training dataset) 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 parameter $\mu_{b,l,s}$ of the voicing model is then estimated as

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

where $v_t(b)$ is the value of the voicing feature.

Examples of the estimated voicing-probabilities $P(v(b) = 1|l, s)$ for HMMs of phonemes /ay/, /v/, and /f/ are depicted in Figure 2. It can be seen that the voicing probabilities for the vowel /ay/ are high over a large number of filterbank channels. Comparing the voiced and unvoiced fricatives /v/ and /f/, respectively, it can be seen that the voicing-probabilities are high for many mixtures (of each state) of the voiced fricative /v/, while close to zero for the unvoiced fricative /f/. Note that the voicing-probabilities at some mixtures of the first and last states of the model /f/ show an increased voicing, which may be due to contextual effects.



(a)

(b)

(c)

Fig. 2. *Examples of the estimated voicing-probabilities for 3 state HMM models of phonemes /ay/, /v/, and /f/ depicted at (a), (b), and (c), respectively. Please note the different scales at each figure.*

3.2 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., $P(\mathbf{y}_t, \mathbf{v}_t|s)$. Assuming 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

$$P(\mathbf{y}_t, \mathbf{v}_t|s) = \sum_{l=1}^{L} P(l|s) \prod_b P(y_t(b)|l, s) P(v_t(b)|l, s) \qquad (6)$$

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 the voicing information about an entire frame, i.e., the voicing information vector $\mathbf{v}$ at a given frame will then consist of a single value indicating whether a frame is voiced or unvoiced. The same equations as above would apply for estimation and incorporation of the voicing-probability. The frame-level voicing information in our experiments was obtained as described at the end of Section 2.

The incorporation of the voicing information as in Eq. 6 may not be effective in noisy conditions due to a possible mismatch between the voicing of the current noisy signal and the trained voicing models. The voicing mismatch may occur as some filter-bank channels which were voiced in a clean signal become unvoiced in a noisy signal due to the effect of noise – an example of this can be observed in Figure 1. The voicing-probability for these channels would then have a small value on the correct voiced model and a large value on any incorrect unvoiced model and as such could negatively affect the recognition. This problem may be dealt with by using the voicing information of the signal during the recognition only when it was estimated as voiced, i.e., marginalising the voicing-probability term in Eq. 6 for features detected as unvoiced. This issue will be demonstrated in the experimental section.

3.3 Transformation of the voicing-probability

During recognition, the voicing-probability may become very small (or zero) on states/mixtures whose trained voicing model have the parameter $\mu_{b,l,s}$ approaching (or equal) to zero or one. This is not desirable, as it can cause the overall probability during the recognition to become largely affected by the voicing-probability. This could be avoided by setting a small minimum value for $P(v(b)|l,s)$. A more elegant solution, also allowing us to easily control the effect of the voicing-probability on the overall probability, may be to employ a sigmoid function to transform the $P(v(b)|l,s)$ for each b to a new value, i.e.,

$$P(v(b)|l,s) = \frac{1}{1 + e^{-\alpha(P(v(b)|l,s)-0.5)}} \tag{7}$$

where α is a constant defining the slope of the function and the value 0.5 gives shift of the function.

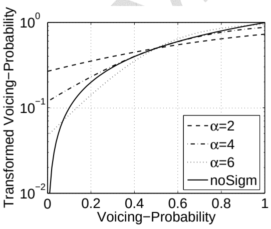


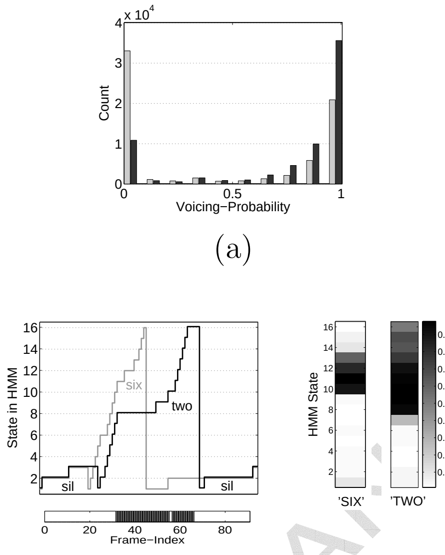
Fig. 3. *Voicing-probability transformation using a sigmoid function with various values of the slope parameter α and no transformation. Note the y-axis is in the log-scale.*

Examples of the voicing-probability transformation with various values for α are depicted on Figure 3. Note that setting α to zero corresponds to models without the incorporation of the voicing-probability. Increasing the value of α increases the effect of the voicing-probability on the overall probability. However, a high value for α may cause the errors in voicing estimation to have a high impact on the overall probability, and consequently affect the recognition performance negatively. An appropriate value for α can be decided based on a small set of experiments on the development data. In our experiments, we have observed that the values of α within the range $[2, 6]$ yielded good recognition results.

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. Frame-level voicing information was considered to simplify the 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 to which the frame is associated 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 in Figure 4(a). It can be seen that when the voicing information is not incorporated (light) 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 4(b) shows an example of the Viterbi-found path for the speech utterance “two” obtained by ASR system without and with incorporated voicing-probability, and the corresponding recognition result obtained as “six” and “two”. Also, the estimated frame-level voicing information of the utterance and the voicing-probabilities of the HMMs are shown at the bottom and the right-side of the figure, respectively. A significant disagreement between the voicing of the model and the 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. 4. *Histogram of state voicing-probabilities associated with voiced frames without (light) and with (dark), using the voicing information (a). Recognition of a speech utterance "two", and state-time path, without (light) and with (dark), 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 evaluations

The experimental evaluation of the proposed model was performed on two speech recognition tasks. First, tests were carried out on the Aurora 2 database for the connected-digit recognition task, as this is currently one of the standard databases used for noisy speech recognition. In order to provide a better analysis of the incorporation of the voicing information, further tests were performed on the TIMIT database for phoneme recognition.

The evaluation was performed first by using standard models trained on clean training data and then by using models that had compensated for the effect of noise. The latter was done in order to determine what performance improvements could be expected by incorporation of the voicing information when idealised noise-compensated models are available since the use of such models may reduce the amount of misalignment between data and models (and thus also voicing misalignment). The missing-feature theory (MFT) was employed as one way of noise compensation. In order to obtain the best (idealised) elimination of the noise effect, the MFT-model with an oracle mask, obtained by full a-priori knowledge of noise, was employed. In Aurora task, models obtained by multi-conditional training were also used as an alternative way of noise compensation.

The experimental evaluation was performed by using an in-house speech recogniser and the Hidden Markov Model Toolkit (HTK) (Young et al., 1999) which was modified to include the missing-feature method and the voicing-probability incorporation.

4.1 Experiments on Aurora 2 database

4.1.1 Database description

The Aurora 2 English language database (Hirsch and Pearce, 2000) was used for speaker-independent connected-digit recognition in noisy conditions. The recognition experiments were performed using speech data from the test set A in the Aurora 2 database which contains 1001 utterances of speech artificially corrupted by four environmental noise types: subway, babble, car, and exhibition hall, each of these at six different SNRs: 20, 15, 10, 5, 0, and -5 dB.

4.1.2 Acoustic modelling

The frequency-filtered logarithm filter-bank energies (Nadeu et al., 2001) (referred here as FF-features) were used as speech feature representation due to their suitability for missing-feature based recognition. It is to be noted that FF-features have previously been shown to yield similar recognition performance as mel-frequency cepstral coefficients (Nadeu et al., 2001). The FF-features were obtained with the following parameter set-up: frames of 32 ms length with a 10 ms shift between the 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 then filtered using the filter $H(z)=z\text{-}z^{-1}$ (Nadeu et al., 2001). A feature vector consisting of 18 elements was obtained (the edge values were excluded). In order to include dynamic spectral information, the first-order delta parameters were added to the static FF-feature vector, resulting in a 36-dimensional feature vector.

The HMMs were trained following the procedures distributed with the Aurora 2 database, and summarised here. Each digit was modelled by a continuous-observation left-to-right HMM with 16 states (no skip allowed), with three and ten Gaussian mixtures for each state in the case of the standard and multi-conditional model, respectively, and with diagonal covariance matrices.

Standard and MFT-based models were trained using the clean speech training set containing 8440 utterances of 55 male and 55 female adult speakers. A multi-conditional model was trained using the multi-conditional training set, which consists of both clean speech and speech corrupted at four different SNRs (20, 15, 10, and 5 dB) by noises from this test-set.

The voicing information was estimated for filter-bank channels as described in Section 2. A FF-feature was then assigned as voiced, i.e., $v(b)$=1, if both of the filter-bank channels involved in the calculation of the FF-feature were voiced, otherwise they were assigned as unvoiced. The voicing models were trained using the clean training data for both the standard and MFT-based model and using the multi-conditional training data for the multi-conditional model. In recognition, models incorporating the voicing information used a sigmoid function with α set to 5 in the case of standard model, and with α set to 3 in the MFT-based and multi-conditional models in all the Aurora 2 experiments.

4.1.3 Experimental analysis of the methods

This section presents an experimental analysis of various ways of incorporating the voicing information as well as examining the effect of errors in voicing estimation. In order to assess the performance of the analysed methods, the experiments are performed on the Aurora 2 speech data corrupted by White noise, as this noise does not contain any pure sinusoidal components.

First, we analyse two ways of incorporating the voicing information during recognition as discussed in Section 3.2 – the model denoted as "VP(0&1)" employs the voicing-probability term in Eq. 6 for both voiced (i.e., 1) and unvoiced (i.e., 0) features and the model denoted as "VP(1)" for only voiced features (i.e., marginalising the voicing-probability term for unvoiced features). The experimental results are presented in Figure 5. It can be seen that the method employing the voicing-probability for both voiced and unvoiced features gives (overall) a significantly lower recognition accuracy than that of employing the voicing-probability only for voiced features. The reason for the low performance obtained when including the voicing-probability term for unvoiced features may be that some of those features may have been voiced on clean speech, thus no longer in agreement with the trained voicing models – this has been discussed in more detail in Section 3.2 and can be also observed in Figure 1. The experimental results presented in the rest of the paper are obtained by using the voicing-probability term only for features detected as voiced and for clarity the notation VP(1) is simplified to VP.

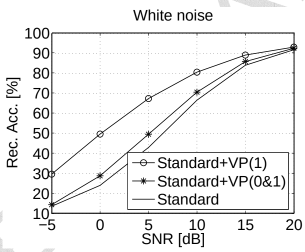


Fig. 5. *Recognition accuracy results obtained by the standard model with the voicing-probability being incorporated for both voiced and unvoiced features and for only voiced features. For comparison, results by the standard model are also included.*

Secondly, we present experiments to demonstrate the effect of possible errors in voicing estimation on the recognition performance when the voicing-probability is incorporated. The experimental results obtained by employing the voicing information estimated on noisy speech are compared to those obtained by employing the oracle voicing information. A filter-bank channel of noisy speech is assigned an oracle label voiced if it was estimated as voiced on corresponding clean speech and its local-SNR is above 0 dB (i.e., speech dominated). The experimental results, presented in Figure 6, show that nearly identical recognition accuracies are obtained when employing the estimated voicing information and the oracle voicing information. These results demonstrate a good performance of the proposed voicing-estimation method presented in Section 2 as the errors it makes have virtually no effect on the recognition accuracy of an ASR system incorporating this voicing information.

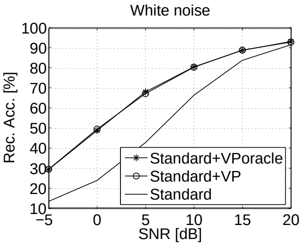


Fig. 6. *Recognition accuracy results obtained by the standard model with incorporated voicing probability employing both the estimated and oracle voicing information. For comparison, results by the standard model are also included.*

Next, we present experiments to compare the effect of incorporating the voicing-probability when feature-level and frame-level voicing information is employed, i.e., voicing information of each filter-bank channel and a single value voicing information of an entire frame, respectively. The experimental results are depicted in Figure 7. It can be seen that the use of a feature-level voicing ("Standard+VP") provides significantly better performance at all SNRs than using the frame-level voicing ("Standard+VPfrm"), which is a consequence of a more detailed modelling of the voicing information.

For comparison, we also incorporated the voicing information in a way employed in some of the previous other research, e.g., Zolnay et al. (2003), in which a frame-level voicing information (representing the level of voicedness of a frame) is appended into the feature vector. The frame-level voicing information we used here represents the proportion of the voiced filter-bank channels in a frame, which provides a measure of voicedness of a frame. The results are depicted in Figure 7 under the notation "Standard+VPoth". It can be seen that the model "Standard+VPoth" obtained some improvements over the standard model, however, these improvements are significantly lower (especially at low SNRs) than those obtained by the proposed "Standard+VP" model.

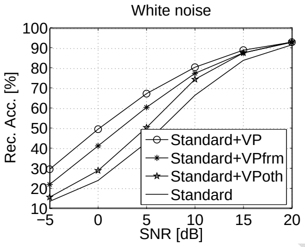


Fig. 7. *Recognition accuracy results obtained by the standard model with incorporated voicing probability employing the feature-level and frame-level voicing information. For comparison, results by the standard model incorporating the voicing information as employed in some other research and standard model alone are also included.*

Last, we discuss the model complexity. In our experimental set-up, the standard model has 108 mean and 108 variance parameters, plus 2 free mixture weights, totalling 218 free parameters per state. The standard model with incorporated feature-level voicing information has an additional 54 parameters, totalling 272 parameters. The multi-conditional model without and with incorporated voicing information has in total 729 and 909 parameters, respectively. Since the incorporation of the voicing information increases the number of parameters, in order to evaluate its effect we performed comparison with the standard model having a similar complexity. In experiments, the standard model with 4 mixtures per state (denoted as "Standard(4mix)") was employed as it has in total 291 parameters, which is similar to 272 parameters needed for the standard model with 3 mixtures and incorporated voicing information. Results are depicted in Figure 8. Comparing the systems of similar complexity, it can be seen that the performance obtained by the model with incorporated voicing information does not come from the increased complexity of the model.

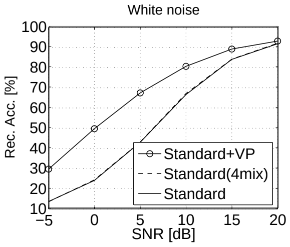


Fig. 8. *Comparison of the recognition accuracy results obtained by models of similar complexity: the standard model with 4 mixtures per state and the standard model with 3 mixtures per state and incorporated voicing information. The standard model with 3 mixtures also included for comparison.*

4.1.4 Experimental results

Experimental results on the standard model

First, the evaluation of the proposed voicing incorporation was performed using a standard model trained on clean data. The results are presented in Figure 9. It can be seen that the incorporation of the voicing-probability provides significant improvements in recognition accuracy. In the case of Babble noise, it was observed that the incorporation of the voicing-probability resulted in a high increase in the number of insertions, and as such a decrease of the recognition accuracy. This is due to the Babble noise being a background speech. This could be dealt with by segmenting the signal into regions of interest and regions of no interest. As this issue is not the subject of this paper, we employed a simple method for segmenting the detected voiced regions into a foreground and background based on their energy, and considered foreground regions to be the speech of interest. The energy-based foreground/background detection was performed by calculating an average of the five highest, E_h, and five lowest, E_l, frame energies within a 500ms segment around the current frame-time and assigning the frame as foreground if its energy was above the value of $E_l + 0.15(E_h - E_l)$. The estimated voicing information was used only if the frame was detected as foreground. Experimental results obtained by employing the voicing information with the foreground detection are denoted by "Standard+VP+FD" in Figure 9. As we can see here, considerable recognition accuracy improvements over the standard model without incorporating the voicing are achieved for Babble noise and also some further improvements for other noisy conditions at low SNRs. The voicing information with the foreground detection is employed in the following experiments in this section.

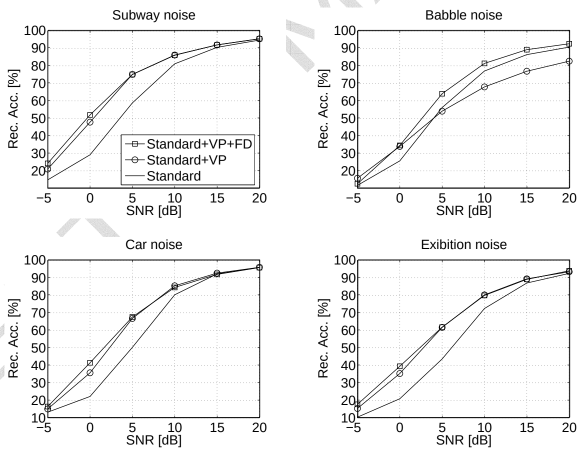


Fig. 9. *Recognition accuracy results obtained by the standard model without and with incorporated voicing probability (without and with incorporated foreground detection).*

The performance of the standard model without and with incorporated voicing is summarised in terms of the average recognition accuracy and error rate reduction calculated over all noises for a given SNR in Table 1 and over SNRs (0–20 dB) for a given noise condition in Table 2. It can be seen that the model incorporating the voicing information provides a large error rate reduction at each SNR, with the highest reductions at SNRs of 10, 5 and 0 dB. Similarly, the proposed model provides large error rate reductions for each noisy condition.

Table 1

Average recognition accuracy results and error rate reduction (ERR) over noisy conditions for the standard model without and with incorporated voicing-probability.

SNR	Avg. Rec. Acc. [%]		ERR
[dB]	Stand+VP	Stand	[%]
20	94.49	93.48	15.45
15	90.57	89.04	13.96
10	82.98	77.73	23.55
5	67.14	52.21	31.24
0	41.83	24.49	22.97
-5	17.76	12.53	5.98
ave (0–20)	75.40	67.39	24.56

Experimental results on noise-compensated models

Next, we performed two sets of evaluations in order to determine whether the incorporation of the voicing information could still provide improvements when used within models that had already suppressed the effect of noise (as employment of a noise compensation would effectively cause the misalignment

Table 2

Average recognition accuracy results and error rate reduction (ERR) over SNRs (0–20 dB) for the standard model without and with incorporated voicing-probability.

Noisy	Avg. Rec. Acc. [%]		ERR
speech	Stand+VP	Stand	[%]
Subway	80.11	70.88	31.71
Babble	72.37	67.20	15.75
Car	76.27	68.18	25.42
Exhib.	72.85	63.30	26.04

of voicing to be less likely).

This is first demonstrated by using a model based on the missing-feature theory (MFT) as a way of noise compensation. In order to obtain the best (idealised) noise compensation, we used the MFT-model with an oracle mask, obtained by full a-priori knowledge of noise. In this model, the static features whose local SNRs are below 0dB were marginalised. The experimental results are presented in Figure 10. It can be seen that the incorporation of the voicing-probability into the MFT-model (“MFT+VP”) results in significant improvements at low SNRs for Subway, Car and Exhibition noisy speech, as well as some improvements on Babble noisy speech. The second noise-compensated model we employed was the model based on the multi-conditional training. The results are presented in Figure 11. Again, the voicing incorporation provides significant improvements at low SNRs.

The performance of the noise-compensated models without and with incorpo-

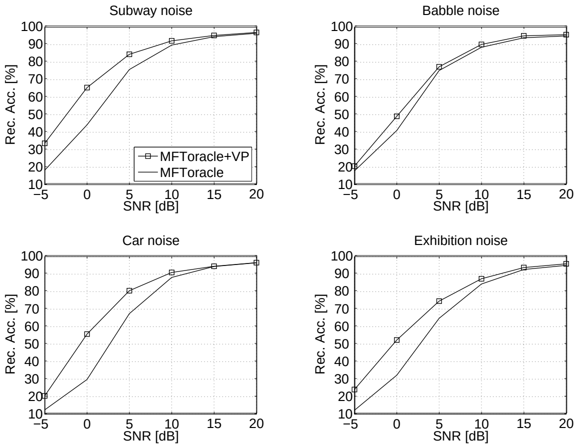


Fig. 10. *Recognition accuracy results obtained by the MFT-model using the oracle mask without and with incorporating the voicing probability.*

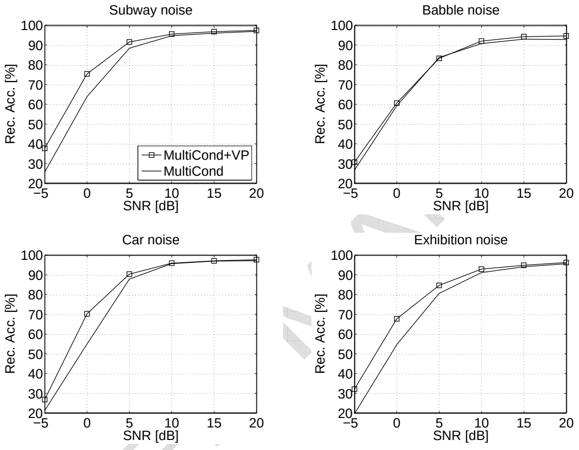


Fig. 11. *Recognition accuracy results obtained by the multi-conditional trained model without and with incorporating the voicing probability.*

rated voicing is summarised in terms of the average recognition accuracy and error rate reduction calculated over all noises for a given SNR in Table 3 and over SNRs (0–20 dB) for a given noisy condition in Table 4. The error rate reductions are large at all SNRs, with the highest reduction at 0 dB SNR, and they are over 24% for all noisy conditions except for the Babble noise. The above results demonstrate that even when the noise effect has already been compensated, the incorporation of the voicing information provides a significant error rate reduction.

Table 3

Average recognition accuracy results and error rate reduction (ERR) over noisy conditions for the MFToracle and MultiCond models without and with incorporated voicing-probability.

SNR	Avg. Rec. Acc. [%]		ERR	Avg. Rec. Acc. [%]		ERR
[dB]	MFTo+VP	MFTo	[%]	MultiC+VP	MultiC	[%]
20	96.31	95.76	13.02	96.78	95.92	20.91
15	94.66	93.90	12.50	96.04	95.28	15.95
10	90.17	87.70	20.09	94.37	93.35	15.31
5	79.26	70.78	29.04	87.74	85.40	16.01
0	55.67	36.82	29.83	68.75	58.40	24.88
-5	24.69	15.21	11.19	32.02	23.46	11.19
ave (0–20)	83.21	76.99	27.03	88.74	85.67	21.42

4.2 Experiments on the TIMIT database

In order to further examine the effect of incorporating the voicing information, we performed experiments for isolated phoneme recognition. Having the start/end information of the phoneme in the signal, these experiments aimed to demonstrate the effect of the voicing incorporation.

Table 4

Average recognition accuracy results and error rate reduction (ERR) over SNRs (0–20 dB) for the MFToracle and MultiCond models without and with incorporated voicing-probability.

Noisy	Avg. Rec. Acc. [%]		ERR	Avg. Rec. Acc. [%]		ERR
speech	MFTo+VP	MFTo	[%]	MultiC+VP	MultiC	[%]
Subway	86.91	80.15	34.04	91.65	88.27	28.77
Babble	81.49	78.75	12.93	85.24	84.18	6.69
Car	83.67	75.26	33.99	90.51	86.76	28.30
Exhibition	80.79	73.81	26.66	87.53	83.46	24.61

4.2.1 Database description

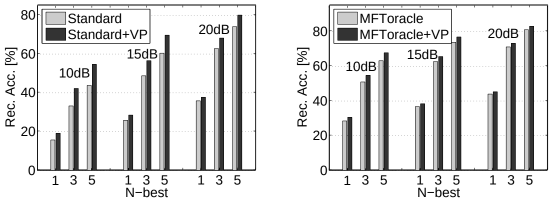
The TIMIT database (Garofolo et al., 1993), downsampled to 8 kHz, was used. The training set comprised speech from all speakers in the TIMIT training set (318 speakers, 5040 utterances). Testing was performed on the TIMIT complete test set (168 speakers, 1344 utterances) corrupted by White noise at SNRs of 20, 15, and 10 dB.

4.2.2 Acoustic modelling

The FF-features as described in Section 4.1.2 were used. The vocabulary consisted of 39 monophones. Each monophone was modelled by a continuous-observation left-to-right HMM with 3 states (no skip allowed), with eight component Gaussian mixtures and diagonal covariance matrices used for each state. The HMMs were trained on clean speech from the training set. No phone language model was employed in recognition experiments, which was considered appropriate for investigating the effect of incorporating voicing information. Note that our baseline recognition system achieved 56% accuracy on clean data which is similar to results for this task presented elsewhere, e.g., (Russell and Jackson, 2005). Models incorporating the voicing information used a sigmoid function with α set to 2.

4.2.3 Experimental results

The experiments were performed using both the standard model and the MFT-model with oracle mask as the noise-compensated model and the obtained results are shown in Figure 12(a) and Figure 12(b), respectively. The results are presented as N-best recognition accuracy (with N equal to 1, 3 and 5), i.e., the correct result being among the first N recognition results. The N-best recognition analysis can provide useful insights for employment of phoneme-level language models, which are often used for re-scoring the recognition results in various areas of speech pattern processing, for instance, in continuous speech recognition for spoken document retrieval (Larson, 2001) and language identification (Zissman and Berkling, 2001). From Figures 12(a) and (b), it can be seen that the incorporation of the voicing information provides considerable recognition accuracy improvements in all cases. With the voicing-probability incorporated, the percentage of the correctly recognised phonemes within the first three and five results is significantly improved when using the standard model and considerably improved when using the noise-compensated model. For instance, the recognition accuracy for N=5 at 10 dB SNR improved from 43.8% to 54.7% and from 63.22% to 67.79% when using the standard model and the noise-compensated model, respectively. Note that the results presented in these two figures are the average over all the phonemes. As the voicing-probability was incorporated only for features being voiced, improvements are not expected for unvoiced phonemes and as such the actual improvements for the voiced phonemes may be considerably higher than those observed in Figure 12. As an example of this, Figure 13 presents the 1-best recognition



(a) (b)

Fig. 12. *The N-best phoneme recognition accuracy obtained by the standard model (a) and MFT-model using oracle mask (b), each without and with incorporating the voicing information. Results on speech corrupted by White noise at SNR of 10, 15 and 20dB.*

accuracy results for each individual phoneme obtained by the standard model at 15 dB SNR. In the figure, phonemes are placed in descending order based on the difference between recognition accuracies obtained by the model with and without incorporating the voicing-probability. It can be seen that the incorporation of the voicing-probability provides large recognition accuracy improvements for many of the voiced phonemes. For instance, phonemes /v/, /y/, /oy/ obtained absolute recognition accuracy improvements of 20.29 %, 12.76 %, and 8.66 %, respectively. It can be observed that some phonemes show a slight decrease in recognition accuracy, which may be due to the context affecting the voicing of the signal.

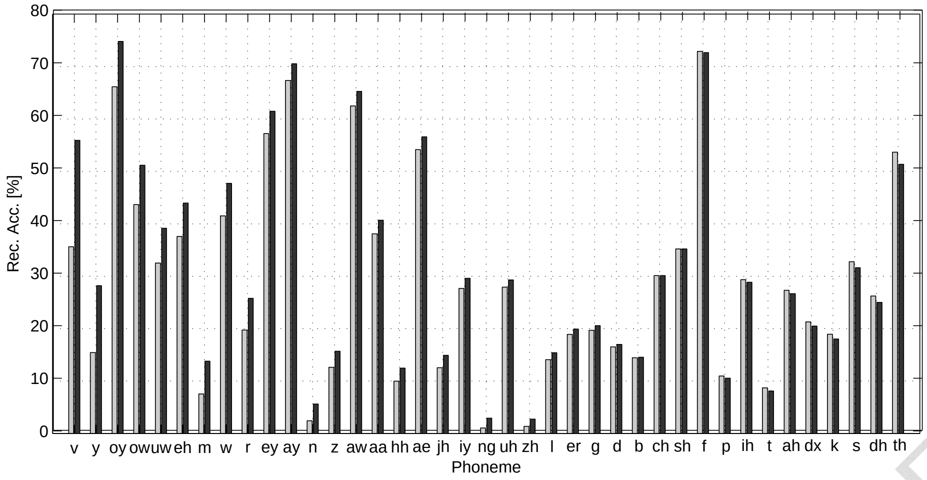


Fig. 13. *Recognition accuracy for individual phonemes obtained by the standard model without and with incorporating the voicing probability for speech corrupted by White noise at 15 dB SNR.*

5 Conclusion

In this paper, we presented a novel model for the incorporation of the voicing information of a speech signal into an HMM-based automatic speech recognition system. The voicing information employed was obtained by a novel method that can provide this information for each filter-bank channel, while requiring no information about the fundamental frequency. The voicing information was modelled by using Bernoulli distribution. A Viterbi-style training procedure for estimation of the voicing-models for each mixture at each HMM state was presented. In the incorporation of the voicing-probability during recognition, the marginalisation of unvoiced voicing information and the use of a sigmoid function to control the contribution of the voicing-probability were proposed. The employment of a frame-level and feature-level voicing information were compared and significant gains by using the feature-level voicing were demonstrated. An experimental evaluation was first performed on noisy speech data from the Aurora 2 database. The effectiveness of the proposed model was demonstrated when employed in both a standard model and in models that had compensated for the effect of noise, missing-feature and multi-conditional training. The experimental evaluations showed that the incorporation of the voicing information within the standard model as well as noise-compensated models provided significant performance improvements in particular at strong noisy conditions. When the voicing information was incorporated, the error rate reduction averaged over all noisy conditions and SNRs was 24.56%, 27.08% and 21.35% for the standard, the MFT-oracle and the multi-conditional training models, respectively. Further experimental evaluations were also performed on phoneme recognition task on a noise-corrupted TIMIT database. These experimental results were analysed as N-best recognition performance. It was shown that considerable performance improvements can be achieved when the voicing information is incorporated in both the standard model and the MFT-oracle noise-compensated model.

6 Acknowledgement

This work was supported by UK EPSRC grants EP/D033659/1 and EP/F036132/1.

References

Beaufays, F., Boies, D., Weintraub, M., Zhu, Q., 2003. Using speech/non-speech detection to bias recognition search on noisy data. ICASSP, Hong-Kong, China I, 424–427.

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.

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

Fant, G., 1960. Acoustic Theory of Speech Production. The Hagues:Mounton.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., 1993. The darpa timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, Philadelphia.

Graciarena, M., Franco, H., Zheng, J., Vergyri, D., Stolcke, A., 2004. Voicing feature integration in SRI's decipher LVCSR system. ICASSP, Montreal, Canada I, 921–924.

Hirsch, H. G., Pearce, D., Sept. 2000. 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.

Huang, H. C.-H., Seide, F., 2000. Pitch tracking and tone features for Mandarin speech recognition. ICASSP, Istanbul, Turkey, 1523–1526.

Ishizuka, K., Nakatani, T., Minami, Y., Miyazaki, N., 2006. Speech feature extraction method using subband-based periodicity and nonperiodicity decomposition. Journal of the Acoustical Society of America 120 (1), 443–452.

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

Jančovič, P., Köküer, M., Jan. 2007a. Estimation of voicing-character of speech spectra based on spectral shape. IEEE Signal Processing Letters 14 (1), 66–69.

Jančovič, P., Köküer, M., 2007b. Incorporating the voicing information into HMM-based automatic speech recognition. IEEE Workshop on Automatic Speech Recognition and Understanding, Kyoto, Japan, 42–46.

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

Kitaoka, N., Yamada, D., Nakagawa, S., 2002. Speaker independent speech recognition using features based on glottal sound source. ICSLP, Denver, USA, 2125–2128.

Larson, M., 2001. Sub-word-based language models for speech recognition: Implications for spoken document retrieval. Proc. of the Workshop on Language Modeling and Information Retrieval, Carnegie Mellon University.

Ljolje, A., 2002. Speech recognition using fundamental frequency and voicing in acoustic modeling. ICSLP, Denver, USA, 2137–2140.

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

Niyogi, P., Ramesh, P., 2003. The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets. Speech Communication 41, 349–367.

O'Shaughnessy, D., Tolba, H., 1999. Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision. ICASSP, Phoenix, Arizona I, 413–416.

Rabiner, L. R., Sambur, M. R., 1976. Some preliminary experiments in the recognition of connected digits. IEEE Trans. on Acoustic, Speech, and Signal Proc. 24 (2), 170–182.

Russell, M. J., Jackson, P. J. B., 2005. A multiple-level linear/linear segmental HMM with a formant-based intermediate layer. Computer Speech and Language 19, 205–225.

Thomson, D. L., Chengalvarayan, R., 2002. The use of voicing features in HMM-based speech recognition. Speech Communication 37, 197–211.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 1999. The HTK Book. V2.2.

Zissman, M. A., Berkling, K. M., 2001. Automatic language identification. Speech Communication 35, 115–124.

Zolnay, A., Schluter, R., Ney, H., 2003. Extraction methods of voicing feature for robust speech recognition. Eurospeech, Geneva, Switzerland, 497–500.