3

Employment of Spectral Voicing Information for Speech and Speaker Recognition in Noisy Conditions

Peter Jančovič and Münevver Köküer
University of Birmingham
United Kingdom

1. Introduction

In this chapter, we describe our recent advances on representation and modelling of speech signals for automatic speech and speaker recognition in noisy conditions. The research is motivated by the need for improvements in these research areas in order the automatic speech and speaker recognition systems could be fully employed in real-world applications which operate often in noisy conditions.

Speech sounds are produced by passing a source-signal through a vocal-tract filter, i.e., different speech sounds are produced when a given vocal-tract filter is excited by different source-signals. In spite of this, the speech representation and modelling in current speech and speaker recognition systems typically include only the information about the vocal-tract filter, which is obtained by estimating the envelope of short-term spectra. The information about the source-signal used in producing speech may be characterised by a voicing character of a speech frame or individual frequency bands and the value of the fundamental frequency (F0). This chapter presents our recent research on estimation of the voicing information of speech spectra in the presence of noise and employment of this information into speech modelling and in missing-feature-based speech/speaker recognition system to improve noise robustness. The chapter is split into three parts.

The first part of the chapter introduces a novel method for estimation of the voicing information of speech spectrum. There have been several methods previously proposed to this problem. In (Griffin & Lim, 1988), the estimation is performed based on the closeness of fit between the original and synthetic spectrum representing harmonics of the fundamental frequency (F0). A similar measure is also used in (McAulay & Quatieri, 1990) to estimate the maximum frequency considered as voiced. In (McCree & Barnwell, 1995), the voicing information of a frequency region was estimated based on the normalised correlation of the time-domain signal around the F0 lag. The author in (Stylianou, 2001) estimates the voicing information of each spectral peak by using a procedure based on a comparison of magnitude values at spectral peaks within the F0 frequency range around the considered peak. The estimation of voicing information was not the primary aim of the above methods, and as such, no performance evaluation was provided. Moreover, the above methods did not consider speech corrupted by noise and required an estimation of the F0, which may be difficult to estimate accurately in noisy speech. Here, the presented method for estimation of

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

the spectral voicing information of speech does not require information about the F0 and is particularly applicable to speech pattern processing. The method is based on calculating a similarity, which we refer to as voicing-distance, between the shape of signal short-term spectrum and the spectrum of the frame-analysis window. To reflect filter-bank (FB) analysis that is typically employed in feature extraction for ASR, the voicing information associated with an FB channel is computed as an average of voicing-distances (within the channel) weighted by corresponding spectral magnitude values. Evaluation of the method is presented in terms of false-rejection and false-acceptance errors.

The second part of the chapter presents an employment of the estimated spectral voicing information within the speech and speaker recognition based on the missing-feature theory (MFT) for improving noise robustness. There have been several different approaches to improve robustness against noise. Assuming availability of some knowledge about the noise, such as spectral characteristics or stochastic model of noise, speech signal can be enhanced prior to its employment in the recogniser, e.g., (Boll, 1979; Vaseghi, 2005; Zou et al., 2007), or noise-compensation techniques can be applied in the feature or model domain to reduce the mismatch between the training and testing data, e.g., (Gales & Young, 1996). Recently, the missing feature theory (MFT) has been used for dealing with noise corruption in speech and speaker recognition, e.g., (Lippmann & Carlson, 1997; Cooke et al. 2001; Drygajlo &. El-Maliki, 1998). In this approach, the feature vector is split into a sub-vector of reliable and unreliable features (considering a binary reliability). The unreliable features are considered to be dominated by noise and thus their effect is eliminated during the recognition, for instance, by marginalising them out. The performance of the MFT method depends critically on the accuracy of the feature reliability estimation. The reliability of spectral-based features can be estimated based on measuring the local signal-to-noise ratio (SNR) (Drygajlo &. El-Maliki, 1998; Renevey & Drygajlo, 2000) or employing a separate classification system (Seltzer et al., 2004). We demonstrate that the employment of the spectral voicing information can play a significant role in the reliability estimation problem. Experimental evaluation is presented for MFT-based speech and speaker recognition and significant recognition accuracy improvements are demonstrated.

The third part of the chapter presents an incorporation of the spectral voicing information to improve speech signal modelling. Up to date, the spectral voicing information of speech has been mainly exploited in the context of speech coding and speech synthesis research. In speech/speaker recognition research, the authors in (Thomson & Chengalvarayan, 2002; Ljolje, 2002; Kitaoka et al., 2002; Zolnay et al., 2003; Graciarena et al., 2004) investigated the use of various measures for estimating the voicing-level of an entire speech frame and appended these voicing features into the feature representation. In addition to voicing features, the information on F0 was employed in (Ljolje, 2002; Kitaoka et al., 2002). In (Thomson & Chengalvarayan, 2002), the effect of including the voicing features under various training procedures was also studied. Experiments in the above papers were performed only on speech signals not corrupted by an additional noise and modest improvements have been reported. In (Jackson et al., 2003), 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

(O'Shaughnessy & Tolba, 1999) 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 (Jančovič & Ming, 2002) 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. The incorporation of the voicing information we present here 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. In the proposed model, having the trained HMMs, with each mixture at each HMM state is associated 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 incorporation of the voicing information is evaluated in a standard model and in a missing-feature model that had compensated for the effect of noise. Experiments are performed on the Aurora 2 database. Experimental results show significant improvements in recognition performance in strong noisy conditions obtained by the models incorporating the voicing information.

2. Estimation of the voicing information of speech spectra

In this section we present a novel method for estimation of the voicing information of speech spectra which was introduced and analysed in (Jančovič & Köküer, 2007a).

2.1 Principle

Speech sounds are produced by passing a source-signal through a vocal-tract filter. The production of voiced speech sounds is associated with vibration of the vocal-folds. Due to this, the source-signal consists of periodic repetition of pulses and its spectrum approximates to a line spectrum consisting of the fundamental frequency and its multiples (referred to as harmonics). As a result of the short-time processing, the short-time Fourier spectrum of a voiced speech segment can be represented as a summation of scaled (in amplitude) and shifted (in frequency) versions of the Fourier transform of the frame-window function. The estimation of the voicing character of a frequency region can then be performed based on comparing the short-time magnitude spectrum of the signal to the spectrum of the frame-window function, which is the principle of the voicing estimation algorithm. Note that this algorithm does not require any information about the fundamental frequency; however, if this information is available it can be incorporated within the algorithm as indicated below.

2.2 Algorithm description

Below are the steps of the algorithm:

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 short-time magnitude-spectrum.

Throughout the chapter we work with signals sampled at Fs=8 kHz, the frame length of 256 samples and the FFT-size of 512 samples.

2. Voicing-distance calculation:

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

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

where k_p is the frequency-index of a spectral peak and L determines the number of components of the spectra at each side around the peak to be compared (the value 2 was used). The spectrum of the signal, $S(k)$, and frame-window, $W(k)$, are normalised to have magnitude value equal to 1 at the peak prior to their use in Eq. 1. The voicing-distances for frequency components around the peak were set to the voicing-distance at the peak, i.e., $vd(k)= vd(k_p)$ for $k\in[k_p\text{-}L,k_p+L]$. Note that if the information about the fundamental frequency is available, the voicing-distance could be calculated at frequency-indices corresponding to multiples of the fundamental frequency instead of the peaks of the spectrum. Also note that the estimate of the fundamental frequency could be obtained based on the minimum cumulated voicing-distance calculated at multiples of considered fundamental frequency values (Jančovič & Köküer, 2007a).

3. Voicing-distance calculation for filter-bank channels:

The voicing-distance for each filter-bank (FB) 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.,

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

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, and $X(b)$ is the overall filter-bank energy value.

4. Post-processing of the voicing-distances:

The voicing-distance obtained from steps (2) and (3) may accidentally become of a low value for an unvoiced region or vice versa. To reduce these errors, we have filtered the voicing-distances by employing 2-D median filters due to their effectiveness in eliminating outliers and simplicity. In our set-up, median filters of size 5x9 and 3x3 (the first number being the number of frames and the second the number of frequency indices) were used to filter the voicing-distances $vd(k)$ and $vd^{fb}(b)$, respectively.

Examples of spectrograms of noisy speech and the corresponding voicing-distances for spectrum and filter-bank channels are depicted in Figure 1.

2.3 Experimental evaluation

This section presents evaluation of the voicing estimation algorithm in terms of false-acceptance (FA) and false-rejection (FR) errors.

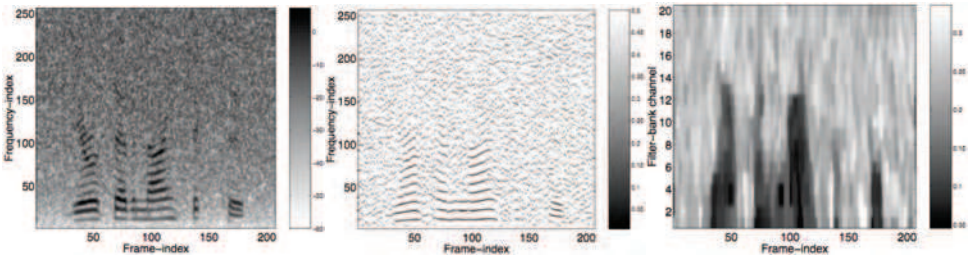

Fig. 1. Spectrogram (left), voicing-distance in the frequency-domain (middle), and in the filter-bank domain (right) of a speech utterance corrupted by white noise at SNR=5dB.

As the true information about the voicing information of FB channels is not available, it is defined based on *a priori* knowledge of clean speech signal and noise; this will be referred to as the "oracle" voicing label. Based on the analysis presented in (Jančovič & Köküer, 2007a), an FB channel of noisy speech is assigned oracle label *voiced* if its corresponding voicing-distance on clean speech is below 0.18 and its local-SNR is above 0 dB, and *unvoiced* otherwise. The decision on the voicing information of an FB channel is made based on a comparison of its corresponding voicing-distance value to a threshold. The experiments were carried out on 2486 digit utterances corrupted by white noise.

The experimental results in terms of FA and FR errors are depicted as a function of the local-SNR in Figure 2. It can be seen that the voicing information can be estimated with a good accuracy; for instance, the FA and FR errors are below 5% when speech signal is corrupted at 10 dB local-SNR and the voicing-distance threshold of 0.21 is used.

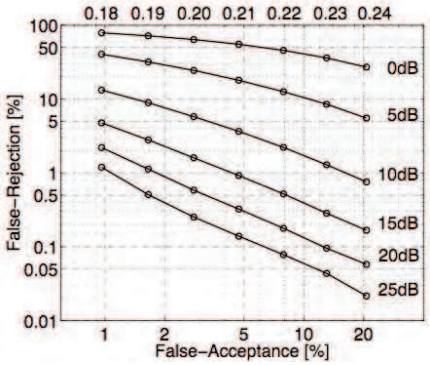

Fig. 2. Performance of the algorithm for estimation of the voicing information of FB channels in terms of FA and FR errors on speech corrupted by white noise. Results presented as a function of the local-SNR and the voicing-distance threshold (depicted above the figure).

3. Employment of the voicing information in the missing-feature-based speech/speaker recognition

The missing-feature theory (MFT) has been successfully employed to improve noise-robustness in automatic speech and speaker recognition systems, e.g., (Cooke, 2001 ; Drygajlo &. El-Maliki, 1998). This approach considers that (in recognition) the feature vector can be split into elements that are not affected (or affected only little) by noise, referred to as

reliable, and elements that are dominated by noise, referred to as *unreliable*. This information is stored in the so-called mask. The unreliable elements are then eliminated during the recognition. In this section we demonstrate that the voicing information of filter-bank channels can provide vital information for estimation of mask employed in MFT-based speech and speaker recognition systems.

3.1 Automatic speech/speaker recognition based on the missing-feature theory

Let $\mathbf{y}_t=(y_t(1), ..., y_t(B))$ denote the feature vector representing the t^{th} frame of the signal and $\mathbf{m}_t=(m_t(1), ..., m_t(B))$ be the corresponding mask vector determining whether an element of $\mathbf{y}_t$ is reliable (equal to 1) or unreliable (equal to 0). Let us consider that speech is modelled by a Hidden Markov Model (HMM) with state output probabilities modelled by a mixture of Gaussian distributions with diagonal covariance matrices. In the marginalisation-based MFT, the observation probability $P(\mathbf{y}_t \mid s,j)$ of the feature vector $\mathbf{y}_t$ at state s and mixture component j is calculated by integrating out the unreliable elements of the feature vector $\mathbf{y}_t$

$$P\left(\mathbf{y}_t \middle| s, j\right) = \prod_{b \in rel} P\left(y_t(b) \middle| s, j\right) \prod_{b \in unrel} \int P\left(y_t(b) \middle| s, j\right) = \prod_{b \in rel} P\left(y_t(b) \middle| s, j\right) \tag{3}$$

We also employed the MFT model within a GMM-based speaker recognition system; Eq. (3) applies also in this case as a GMM can be seen as a 1-state HMM.

In order to apply the MFT marginalisation model, the noise-corruption needs to be localised into a subset of features. This makes the standard full-band cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980), which are currently the most widely used parameterisation of speech, unsuitable as the application of DCT over the entire vector of logarithm filter-bank energies (logFBEs) will cause any corruption localised in the frequency-domain to become spread over all cepstral coefficients. The logFBEs may be employed in the MFT model, however, they suffer from a high correlation between the features, which makes the diagonal covariance matrix modelling not appropriate. The parameterisations often used in MFT model are the sub-band cepstral coefficients, e.g., (Bourlard & Dupont, 1996), and frequency-filtered logFBEs (FF-logFBEs), e.g., (Jančovič & Ming, 2001). The FF-logFBEs, which are obtained by applying a (short) FIR filter over the frequency dimension of the logFBEs, were employed in this paper. These features in standard recognition system have been shown to obtain similar performance as the standard full-band cepstral coefficients (Nadeu et al., 2001), while having the advantage of retaining the noise-corruption localised.

3.2 Mask estimation for the MFT-based ASR

The performance of the MFT-based recognition system depends critically on the quality of the mask. The mask estimation is a complex task and the design of a method for this task is not of our main focus here, rather, we aim to demonstrate that the voicing estimation method can provide useful information to be employed for this task. As such, we are here concerned only with mask estimation for speech detected as voiced in clean data.

We have demonstrated in (Jančovič & Köküer, 2007a) that the voicing-distance $vd^{fb}(b)$ is related to the local-SNR of a voiced FB-channel corrupted by noise. Based on this, the voicing-distance can be used to define the *voicing mask* as

$$m_t^{voic}(b) = 1 \qquad \text{if} \qquad vd_t^{fb}(b) < \beta \tag{4}$$

where the threshold β was set to 0.21. In order to evaluate the quality of the estimated voicing mask, i.e., the effect of errors in the voicing information estimation on the recognition performance, we defined *oracle voicing mask* for an FB-channel as 1 if and only if the channel is estimated as voiced on clean data and its oracle mask (defined as below) is 1 on noisy data.

The so-called *oracle mask* is derived based on full a-priori knowledge of the noise and clean speech signal. The use of this mask gives an upper bound performance and thus it indicates the quality of any estimated mask. We used a-priori SNR to construct the oracle mask as

$$m_t^{\text{oracle}}(b) = 1 \qquad \text{if} \qquad 10\log(X_t(b) / N_t(b)) > \gamma \tag{5}$$

where $X_t(b)$ and $N_t(b)$ are the filter bank energy of clean speech and noise, respectively. The threshold γ was set to -6dB as it was shown to provide a good performance in our experiments.

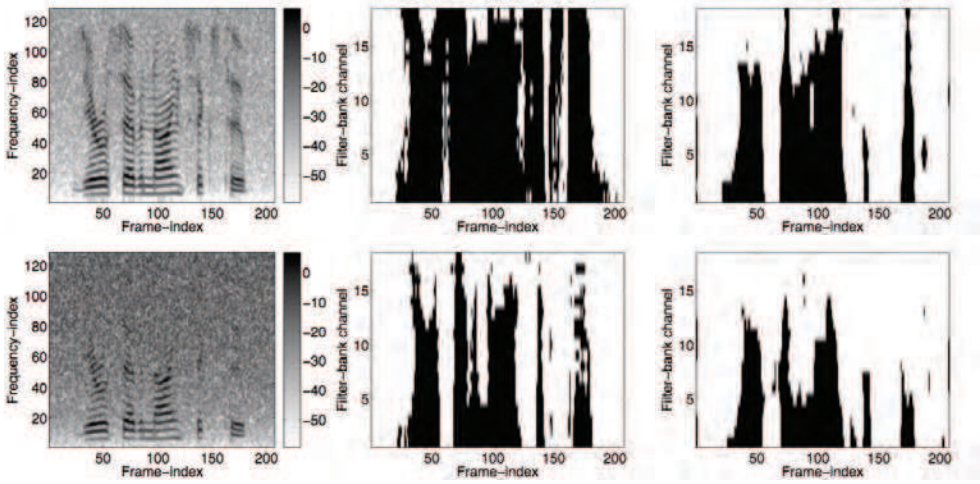

Fig. 3. Spectrograms (a), oracle masks (b), and voicing masks (c) for a speech utterance corrupted by white noise at the SNR=20dB (upper) and SNR=5dB (lower). Mask value equal to 1 and 0 depicted in black and white, respectively.

Figure 3 depicts spectrograms of a speech utterance corrupted by noise at two SNRs and the corresponding oracle and voicing masks. It can be seen that when the noise is strong, the voicing mask is relatively similar to the oracle mask as it is mainly the voiced regions (that are of a higher energy) that are affected only little by noise.

3.3 Experimental evaluations on speech recognition

Experimental evaluations were performed on the Aurora 2 English language database (Hirsch & Pearce, 2000). This database was designed for speaker-independent recognition of digit sequences in noisy conditions. The test set A from the database was used for recognition experiments. This test set consists of four sub-sets, each sub-set contains 1001 utterances of clean speech and noisy speech created by artificially adding to clean speech

one of 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. The clean speech training set, containing of 8440 utterances of 55 male and 55 female adult speakers, was used for training the parameters of HMMs.

The frequency-filtered logarithm filter-bank energies (FF-logFBEs) (Nadeu et al., 2001) were used as speech feature representation, due to their suitability for MFT-based recognition as discussed earlier. Note that the FF-logFBEs achieve similar performance (in average) as standard MFCCs. The FF-logFBEs were obtained with the following parameter set-up: frames of 32 ms length with a shift 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 $H(z)=z-z^{-1}$. 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.

The HMMs were trained following the procedures distributed with the Aurora 2 database. Each digit was modelled by a continuous-observation left-to-right HMM with 16 states (no skip allowed) and three Gaussian mixture components with diagonal covariance matrices for each state. During recognition, the MFT-based system marginalised static features according to the mask employed, and used all the delta features.

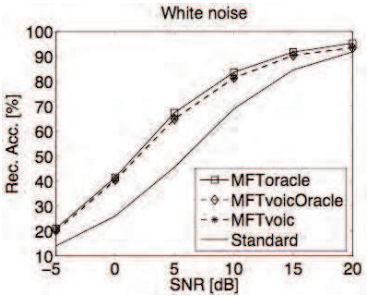

Fig. 4. Recognition accuracy results obtained by the MFT-based speech recognition system employing the voicing and oracle masks.

Experimental results are first presented in Figure 4 for speech corrupted by white noise (as this noise is considered to contain no voiced components) to evaluate the quality of the estimated voicing mask. It can be observed that the recognition accuracies achieved by the MFT-based recogniser employing the estimated voicing mask (MFTvoic) and the oracle voicing mask (MFTvoicOracle) are nearly identical. This indicates that the errors made in the voicing estimation have nearly no effect on the recognition performance of the MFT-based system in the given recognition task. It can also be seen that the MFT-based system using the voicing mask provides recognition performance significantly higher than that of the standard recognition system and indeed very close to using the oracle mask (MFToracle), i.e., the abandonment of uncorrupted unvoiced features did not harm significantly the recognition accuracy in the given task.

Results of experiments on the Aurora 2 noisy speech data are presented in Figure 5. It can be seen that employing the voicing mask in the MFT-based system provides significant

performance improvements over the standard system for most of the noisy conditions. The performance is similar (or lower) than that of the standard system at 20 dB SNR which is due to eliminating the uncorrupted unvoiced features. The MFT-based system employing the estimated voicing mask achieves less improvement in the case of Babble noise because this noise contain voiced components and thus the voicing mask captures the location of both the voiced speech regions as well as the voiced noise regions. There may be various ways to deal with the voiced noise situation. For instance, a simple way may be to consider that the speech is of a higher energy than noise and as such use only the higher energetic voiced regions. Also, signal separation algorithms may be employed, for instance, we have demonstrated in (Jančovič & Köküer, 2007b) that the sinusoidal model can be successfully used to separate two harmonic or speech signals.

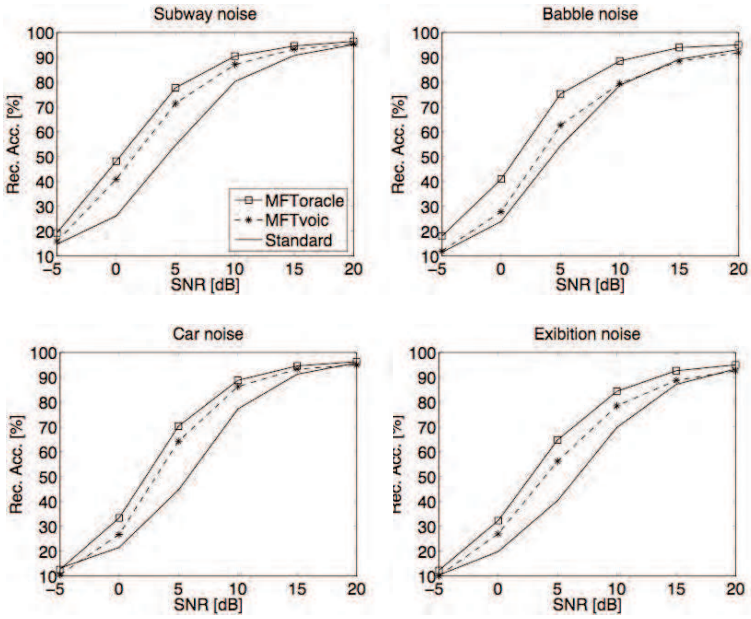

Fig. 5. Recognition accuracy results obtained by the MFT-based speech recognition system employing the voicing and oracle masks.

3.4 Experimental evaluations on speaker recognition

Experiments were performed on the TIMIT database (Garofolo et al., 1993), down sampled to 8 kHz. Hundred speakers (consisting of 64 male and 36 female) from the test subset were selected in an alphabetical order. The training data for each speaker comprised of eight sentences ('si' and 'sx'). The testing was performed using two ('sa') sentences corrupted by Gaussian white noise and Subway noise from the Aurora 2 database, at global SNRs equal to 20, 15, 10, and 5 dB, respectively. The speech feature representation was the same as used in the speech recognition experiments in the previous section. The speaker recognition system was based on Gaussian mixture model (GMM) with 32 mixture-components for each speaker, which was constructed using the HTK software (Young et al., 1999). The GMM for

each speaker was obtained by using the MAP adaptation of a general speech model, which was obtained from the training data from all speakers.
Experimental results are depicted in Figure 6. We can see that the results are of a similar trend as in the case of speech recognition. The use of the estimated voicing mask gives results close to those obtained using the oracle voicing mask. These results are significantly higher than those of the standard model and reasonably close those obtained by the MFT model using the oracle mask which assumes full a-priori knowledge of the noise. These results therefore demonstrate the effectiveness of the estimated voicing mask.

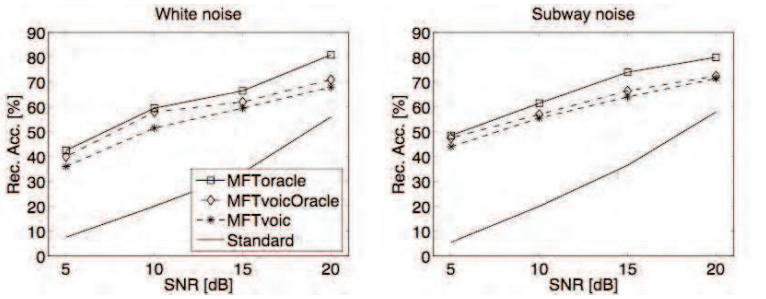

Fig. 6. Recognition accuracy results obtained by the MFT-based speaker recognition system employing the voicing and oracle masks.

4. Modelling the voicing information in the HMM-based ASR

This section presents an employment of the voicing information in an HMM-based ASR system in order to model the properties of the source-signal used to produce the speech. The modelling of the voicing information could be performed by appending the voicing features into the spectral feature vector and using the standard HMM training on the extended feature representation. We have here adopted an alternative approach, in which a separate training procedure for estimation of the voicing models is performed after the HMMs were trained on the spectral features. The presented method was introduced in (Jančovič & Köküer, 2007c).

4.1 Incorporation of the voicing information during recognition

Let $\mathbf{y}_t=(\mathbf{y}_t(1), ..., \mathbf{y}_t(B))$ denote the spectral-feature vector and $\mathbf{v}_t=(\mathbf{v}_t(1), ..., \mathbf{v}_t(B))$ the corresponding voicing vector at the frame-time t, where B is the number of FB channels.
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., $P(\mathbf{y}_t \mid s)$, is replaced by calculating the joint probability of the spectral feature vector and the voicing vector $\mathbf{v}_t$, i.e., $P(\mathbf{y}_t,\mathbf{v}_t \mid s)$. Considering that all spectral features and voicing features are independent of one another, using J mixture densities the $P(\mathbf{y}_t,\mathbf{v}_t \mid s)$ is calculated in the proposed model as

$$P\left(\mathbf{y}_t, \mathbf{v}_t \middle| s\right) = \sum_{j=1}^{J} P\left(j \middle| s\right) \prod_{b} P\left(y_t(b) \middle| s, j\right) P\left(v_t(b) \middle| s, j\right) \tag{6}$$

where $P(j \mid s)$ is the weight of the j^{th} mixture component, and $P(y_t(b) \mid s,j)$ and $P(v_t(b) \mid s,j)$ are the probability of the b^{th} spectral feature and voicing feature, respectively, given state s and

mixture j. Note that the voicing-probability term in Eq.(6) was used only when the feature was detected as voiced, i.e., the term was marginalised for features detected as unvoiced.

4.2 Estimation of the voicing-probability for HMM states

The estimation of the voicing-probability $P(\mathbf{v}\,|\,s,j)$ at each HMM state s and mixture j can be performed using the training data-set by Baum-Welch or Viterbi-style training procedure; the latter was used here.

Given a speech utterance, for each frame t we have the spectral-feature vector $\mathbf{y}_t$ and corresponding 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 j (at the state s) have generated the feature vector $\mathbf{y}_t$ is then calculated as

$$\mathrm{P}\left(\mathrm{j}\middle|\mathbf{y}_\mathrm{t},\mathrm{s}\right)=\frac{\mathrm{P}\left(\mathrm{y_t}\middle|\mathrm{s},\mathrm{j}\right)\,\mathrm{P}\left(\mathrm{j}\middle|\mathrm{s}\right)}{\sum_{\mathrm{j'}}\mathrm{P}\left(\mathrm{y_t}\middle|\mathrm{s},\mathrm{j'}\right)\,\mathrm{P}\left(\mathrm{j'}\middle|\mathrm{s}\right)} \tag{7}$$

where the mixture-weight $P(j\,|\,s)$ and the probability density function of the spectral features used to calculate the $P(\mathbf{y}_t\,|\,s,j)$ are obtained as an outcome of the standard HMM training. For each mixture j and HMM state s, the posterior probabilities $P(j\,|\,\mathbf{y}_t,s)$ for all $\mathbf{y}_t$'s associated with the state s are collected (over the entire training data-set) together with the corresponding voicing vectors $\mathbf{v}_t$'s. The voicing-probability of the b^{th} feature can then be obtained as

$$\mathrm{P}\left(\mathrm{v(b)}=\mathrm{a}\middle|\mathrm{s},\mathrm{j}\right)=\frac{\sum_{\mathrm{t:y_t}\in\mathrm{s}}\mathrm{P}\left(\mathrm{j}\middle|\mathrm{y_t},\mathrm{s}\right)\cdot\delta\left(\mathrm{v_t(b)},\mathrm{a}\right)}{\sum_{\mathrm{t:y_t}\in\mathrm{s}}\mathrm{P}\left(\mathrm{j}\middle|\mathrm{y_t},\mathrm{s}\right)} \tag{8}$$

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

Examples of the estimated voicing-probabilities for HMMs of digits are depicted in Figure 7. It can be seen that, for instance, the first five states of the word 'seven' have the probability of being voiced close to zero over the entire frequency range, which is likely to correspond to the unvoiced phoneme /s/.

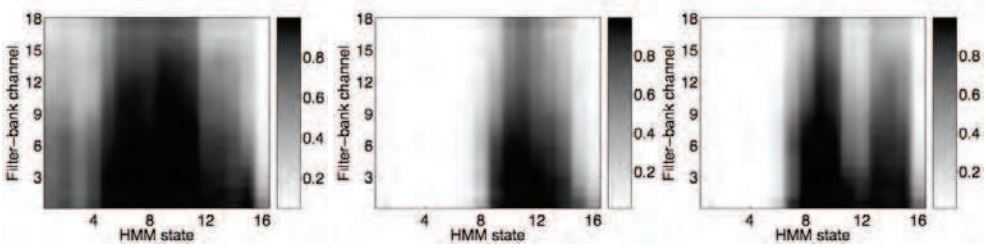

Fig. 7. Examples of the estimated voicing-probabilities for a 16 state HMM models of words 'one' (left), 'two' (middle), and 'seven' (right).

The estimated voicing-probability $P(v(b)\,|\,s,j)$ becomes zero when all features associated with the state are only voiced or unvoiced. This is not desirable, because it can cause the overall probability in Eq.(6) to become zero during the recognition. This could be avoided by setting a small minimum value for $P(v(b)\,|\,s,j)$. A more elegant solution that would also allow us to easily control the effect of the voicing-probability on the overall probability may be to employ a sigmoid function to transform the $P(v(b)\,|\,s,j)$ for each b to a new value, i.e.,

$$P\left(v(b)\middle|s,j\right)=\frac{1}{1+e^{-\alpha\left(P(v(b)|s,j)-0.5\right)}} \tag{9}$$

where α is a constant defining the slope of the function and the value 0.5 gives the shift of the function. Examples of the voicing-probability transformation with various values for α are depicted in Figure 8. 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. The value 1.5 is used for all experiments.

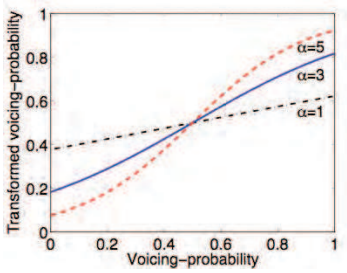

Fig. 8. Sigmoid function with various values of the slope parameter α employed for transformation of the voicing-probability.

4.3 Experimental evaluation

The experiments were performed on the Aurora 2 database. The experimental set-up is the same as described in Section 3.3.

The evaluation of the proposed model for voicing incorporation is first performed using a standard model trained on clean data. Results are presented in Figure 9. It can be seen that the incorporation of the voicing-probability provides significant recognition accuracy improvements 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 of a voiced character. This could be improved by employing a speech-of-interest detection similar as discussed earlier in Section 3.3.

Next, evaluations were performed on a model that had compensated for the effect of noise – these experiments were conducted in order to determine whether the incorporation of the voicing information could still provide improvements (as employment of noise compensation would effectively decrease the amount of misalignment of the voicing information between the signal and models). For this, the marginalisation-based MFT model was used. In order to obtain the best (idealised) noise compensation, this model employs the oracle mask, obtained based on the full a-priori knowledge of the noise. Experimental results are presented in Figure 10. It can be seen 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 and Exhibition noise is, similarly as in the

standard model discussed earlier, due to the voiced character of the noise. It can be seen that the incorporation of the voicing-probability provides significant recognition accuracy improvements at low SNRs, even the noise effect had already been compensated.

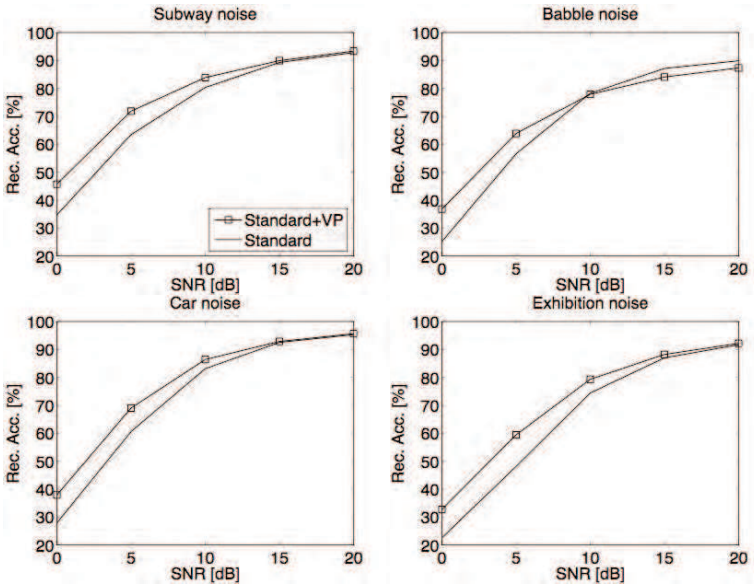

Fig. 9. Recognition accuracy results obtained by the standard ASR system without and with incorporating the voicing-probability.

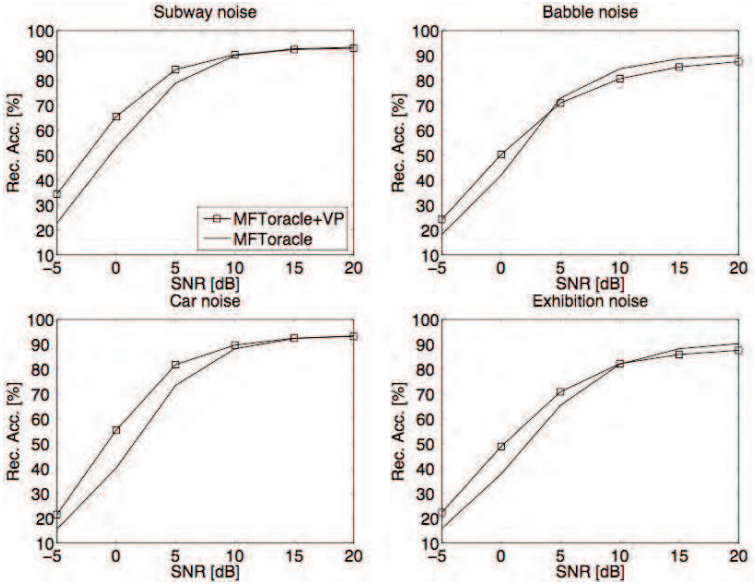

Fig. 10. Recognition accuracy results obtained by the MFT-based ASR system employing the oracle mask without and with incorporating the voicing-probability.

7. Conclusion

This chapter described our recent research on representation and modelling of speech signals for automatic speech and speaker recognition in noisy conditions. The chapter consisted of three parts. In the first part, we presented a novel method for estimation of the voicing information of speech spectra in the presence of noise. The presented method is based on calculating a similarity between the shape of signal short-term spectrum and the spectrum of the frame-analysis window. It does not require information about the F0 and is particularly applicable to speech pattern processing. Evaluation of the method was presented in terms of false-rejection and false-acceptance errors and good performance was demonstrated in noisy conditions. The second part of the chapter presented an employment of the voicing information into the missing-feature-based speech and speaker recognition systems to improve noise robustness. In particular, we were concerned with the mask estimation problem for voiced speech. It was demonstrated that the MFT-based recognition system employing the estimated spectral voicing information as a mask obtained results very similar to those of employing the oracle voicing information obtained based on full a-priori knowledge of noise. The achieved results showed significant recognition accuracy improvements over the standard recognition system. The third part of the chapter presented an incorporation of the spectral voicing information to improve modelling of speech signals in application to speech recognition in noisy conditions. The voicing-information was incorporated within an HMM-based statistical framework in the back-end of the ASR system. In the proposed model, a voicing-probability was estimated for each mixture at each HMM state and it served as a penalty during the recognition for those mixtures/states whose voicing information did not correspond to the voicing information of the signal. The evaluation was performed in the standard model and in the missing-feature model that had compensated for the effect of noise and experimental results demonstrated significant recognition accuracy improvements in strong noisy conditions obtained by the models incorporating the voicing information.

8. References

Boll, S.F. (1979). Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. on Acoustic, Speech, and Signal Proc.*, Vol. 27, No. 2, pp. 113–120, Apr. 1979.

Bourlard, H. & Dupont, S. (1996). A new ASR approach based on independent processing and recombination of partial frequency bands, *Proceedings of ICSLP*, Philadelphia, USA, 1996.

Cooke, M.; Green, P.; Josifovski, L. & Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication,* Vol.34, No. 3, 2001, pp.267–285.

Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Trans. on Acoustic, Speech, and Signal Proc.*, Vol. 28, No. 4, 1980, pp. 357–366.

Drygajlo A. &. El-Maliki, M. (1998). Speaker verification in noisy environment with combined spectral subtraction and missing data theory, *Proceedings of ICASSP*, Seattle, WA, Vol. I, pp. 121–124, 1998.

Gales M.J.F. & Young, S.J. (1996). Robust continuous speech recognition using parallel model combination, *IEEE Trans. on Speech and Audio Proc.,* Vol. 4, pp. 352–359, 1996.

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, *Proceedings of ICASSP,* Montreal, Canada, pp. 921–924, 2004.

Griffin, D. & Lim, J. (1988). Multiband-excitation vocoder, *IEEE Trans. On Acoustic, Speech, and Signal Proc.*, Vol. 36, Feb. 1988, pp. 236–243.

Hirsch, H. & Pearce, D. (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, 2000.

Jackson, P.; Moreno, D.; Russell, M. & Hernando, J. (2003). Covariation and weighting of harmonically decomposed streams for ASR, *Proceedings of Eurospeech*, Geneva, Switzerland, pp. 2321–2324, 2003.

Jančovič, P. & Köküer, M. (2007a). Estimation of voicing-character of speech spectra based on spectral shape, *IEEE Signal Processing Letters,* Vol. 14, No. 1, 2007, pp. 66–69.

Jančovič, P. & Köküer, M. (2007b). Separation of harmonic and speech signals using sinusoidal modeling, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* New York, USA, Oct. 21-24, 2007.

Jančovič, P. & Köküer, M. (2007c). Incorporating the voicing information into HMM-based automatic speech recognition, *IEEE Workshop on Automatic Speech Recognition and Understanding,* Kyoto, Japan, pp. 42–46, Dec. 6-13, 2007.

Jančovič, P. & Ming, J. (2001). A multi-band approach based on the probabilistic union model and frequency-filtering features for robust speech recognition, *Proceedings of Eurospeech*, Aalborg, Denmark, pp. 1111–1114, 2001.

Jančovič, P. & Ming, J. (2002). Combining the union model and missing feature method to improve noise robustness in ASR, *Proceedings of ICASSP*, Orlando, Florida, pp. 69–72, 2002.

Kitaoka, N.; Yamada, D. & Nakagawa, S. (2002). Speaker independent speech recognition using features based on glottal sound source, *Proceedings of ICSLP*, Denver, USA, pp. 2125–2128, 2002.

Lippmann, R.P. & Carlson, B.A. (1997). Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise, *Proceedings of Eurospeech*, Rhodes, Greece, pp. 37–40, 1997.

Ljolje, A. (2002). Speech recognition using fundamental frequency and voicing in acoustic modeling, *Proceedings of ICSLP*, Denver, USA, pp. 2137–2140, 2002.

McAulay, R. J. & Quatieri, T. F. (1990). Pitch estimation and voicing detection based on a sinusoidal speech model, *Proceedings of ICASSP*, pp. 249–252, 1990.

McCree, A.V. & Barnwell, T.P. (1995). A mixed excitation LPC vocoder model for low bit rate speech coding, *IEEE Trans. on Acoustic, Speech, and Signal Proc.,* Vol. 3, No. 4, July 1995, pp. 242–250.

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

O'Shaughnessy, D. & Tolba, H. (1999). Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision, *Proceedings of ICASSP*, Phoenix, Arizona, pp. 413–416, 1999.

Renevey, P. & Drygajlo, A. (2000). Statistical estimation of unreliable features for robust speech recognition, *Proceedings of ICASSP*, Istanbul, Turkey, pp. 1731–1734, 2000.

Seltzer, M.L.; Raj, B. & Stern, R.M. (2004). A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition, *Speech Communication,* Vol. 43, pp. 379–393, 2004.

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis, *IEEE Trans. on Speech and Audio Proc.,* Vol. 9, No.1, Jan. 2001, pp. 21–29.

Thomson, D. & Chengalvarayan, R. (2002). The use of voicing features in HMM-based speech recognition, *Speech Communication*, Vol. 37, 2002, pp. 197–211.

Vaseghi, S.V. (2005). *Advanced Digital Signal Processing and Noise Reduction*, John Wiley & Sons, 2005.

Young, S.; Kershaw, D.; Odell, J.; Ollason, D.; Valtchev, V. & Woodland, P. (1999). *The HTK Book*. V2.2.

Zolnay, A.; Schluter, R. & Ney, H. (2003). Extraction methods of voicing feature for robust speech recognition, *Proceedings of Eurospeech*, Geneva, Switzerland, pp. 497–500, 2003.

Zou, X.; Jančovič, P.; Liu, J. & Köküer, M. (2007). ICA-based MAP algorithm for speech signal enhancement, *Proceedings of ICASSP*, Honolulu, Hawaii, Vol. IV, pp. 569-572, April 14-20, 2007.

Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9
Hard cover, 550 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Peter Jančovič and Münevver Köküer (2008). Employment of Spectral Voicing Information for Speech and Speaker Recognition in Noisy Conditions, Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9, InTech, Available from: http://www.intechopen.com/books/speech_recognition/employment_of_spectral_voicing_information_for_speech_and_speaker_recognition_in_noisy_conditions

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821