EMPLOYMENT OF VOICING INFORMATION OF SPEECH SPECTRA FOR NOISE-ROBUST SPEAKER IDENTIFICATION

Peter Jančovič and Münevver Köküer

Electronic, Electrical & Computer Engineering, University of Birmingham, Birmingham, UK
{p.jancovic, m.kokuer}@bham.ac.uk

ABSTRACT

This paper presents a novel method for voicing information estimation of individual frequency-regions of speech spectra and its employment in a text-independent speaker identification system. The voicing information is incorporated to the system in a form of a mask in a marginalization-based missing-feature model. Experiments were performed on speech data from the TIMIT database corrupted by stationary and real-world noises. The obtained results show that using the proposed voicing estimation method provides performance close to using oracle voicing information. The combination of the voicing information mask with a noise-estimate mask showed further improvement in the identification accuracy and achieved performance close to the oracle mask obtained using full a-priori knowledge of noise.

1. INTRODUCTION

The performance of automatic speaker/speech recognition systems degrades rapidly when speech signal is corrupted by a background acoustical noise. There have been several different ways of improving noise robustness. Speech signal can be enhanced prior to its employment in the recognizer by techniques such as spectral subtraction [1], Wiener filtering, e.g., [2], or exploiting higher-order statistics [3]. Assuming availability of some knowledge about the noise, such as spectral characteristics or stochastic model of noise, noise-compensation techniques, e.g., [4], can be applied in the feature or model domain to reduce the mismatch between the training and testing data.

Recently, the missing feature theory (MFT) has been used for dealing with noise corruption in speech and speaker recognition, e.g., [5] [6] [7]. In this approach, each element of the feature vector is assigned during recognition a label of its reliability. When using binary reliability values, the feature vector is split into a sub-vector of reliable and unreliable features. The unreliable features are either imputed or marginalized out. The performance of the MFT method depends critically on the accuracy of feature reliability estimation. The reliability of filter-bank channels can be estimated based on measuring the local signal-to-noise ratio (SNR) [7] [8]. Recently, a Bayesian classifier, which in addition to the local SNR estimation exploits some characteristic properties of speech signals, was proposed for this estimation [9].

It is generally recognized that some parts of speech signal provide more discriminative information for speaker recognition than others. This has been demonstrated by several studies presenting the speaker discrimination properties of individual phonemes, e.g., [10]. These studies concluded that nasals and vowels provide the best performance for speaker identification in clean speech. Similar analysis presented in [11] showed that in speech contaminated by noise, the unvoiced speech phonemes exhibit significantly lower speaker discriminating properties than voiced speech phonemes, which can be attributed to their lower energy and hence being prone to be affected by noise.

We have recently introduced a novel method that enables to estimate the voicing information (i.e., voiced/unvoiced) of individual frequency-regions of speech spectra [12]. It has been demonstrated that the voiced frequency-regions corrupted by White noise at 10dB local SNR can be detected at below 5% false acceptance and false rejection rate. In this paper, we present an employment of the voicing information obtained by this method in a missing-feature based text-independent speaker identification system. It is shown that the employment of the voicing information of individual frequency-regions provides substantial performance gains over the voicing information of entire frame. The MFT model using the estimated voicing mask of individual frequency-regions is found to achieve very close performance to the one obtained by using oracle voicing mask. We also analysed the effect of marginalizing the delta features on the identification accuracy. Finally, the estimated voicing mask is combined with a mask obtained based on the noise-estimate in order to utilize the reliable features among the unvoiced features. The experimental results show that using the combined mask improves over the individual voicing and noise-estimate masks, and indeed provide performance close to using the oracle mask constructed based on full a-priori knowledge about the noise. All the presented experiments are performed on the TIMIT database corrupted by stationary and non-stationary noises at various SNR levels.

2. ESTIMATING THE VOICING INFORMATION OF FILTER-BANK CHANNELS

This section presents a summary of steps of the algorithm employed for estimation of the voicing information of a signal for each filter-bank channel. This algorithm, introduced in [12] where various experimental evaluations and further analysis were also presented, exploits the quasi-periodicity of voiced speech signals and the effect of short-time processing – due to these, the shape of short-time magnitude spectra of voiced speech around each harmonic frequency should follow approximately the shape of the magnitude spectra of the frame analysis window. Note that this method does not require any information about the fundamental frequency.

Below are the steps of the method:

1) Short-time magnitude-spectra calculation:

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

2) Voicing-distance calculation:

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

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

where k_p is frequency-index of a spectral peak and M determines the number of components of the spectra at each side around the peak to be compared. The spectra of the signal, $S(k)$, and frame-window, $W(k)$, are normalized to have magnitude value equal to 1 at the peak prior to their use in Eq. 1. The range of practical values for the FFT-size (i.e., frame length plus appended zeros) and M can be found in [12]. Here we used FFT-size of 512 points and $M = 3$.

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.,

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

where $G_b(k)$ is the frequency-response of the filter-bank channel b, and k_b and K_b are the lowest frequency-component and number of components of the frequency response, respectively. The $Y(b) = \sum_{k=k_b}^{k_b+K_b-1} G_b(k)|S(k)|^2$, i.e., the overall filter-bank energy value. The Eq. 2 requires voicing-distance values for each frequency component. These can be estimated, for instance, by using a linear interpolation between voicing-distance values corresponding to adjacent peaks.

4) Postprocessing of the voicing-distances: The voicing-distance obtained from Eq. 1 and Eq. 2 may accidentally become of a low value for a unvoiced region or vice versa. This can be improved by filtering of voicing-distance values. Based on the results presented in [12], the filtering was performed on both the interpolated $vd(k)$ and $vd^{fb}(b)$ values, and 2D median filters of size 5×9 and 3×3 (the first number being the number of frames), respectively, were employed.

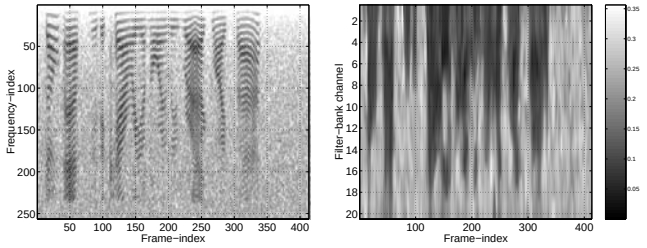
An example of a spectrogram of noisy speech and the corresponding voicing distances for filter-bank channels are depicted on Figure 1.

3. MISSING-FEATURE GMM-BASED SPEAKER IDENTIFICATION

3.1 Marginalisation-based missing-feature model

The missing feature theory has been successfully applied to automatic speech and speaker recognition, e.g., [6] [7]. In this paper, we study the employment of the method for estimation of the voicing information presented in the previous section within the MFT model for a text-independent speaker identification.

We consider that each speaker is modelled by a Gaussian mixture model (GMM) whose parameters are obtained using all features from the clean training data. In recognition, it



(a) (b)

Figure 1: *An example of spectrogram of speech utterance corrupted by White noise at 15dB (a) and corresponding voicing distance for filter-bank channels (b).*

is considered that a feature vector consists of elements that are not affected (or affected only little) by the noise, referred to as *reliable*, and elements that are strongly corrupted by noise, referred to as *unreliable*. Considering Gaussian densities with a diagonal covariance matrix, the probability of the feature vector $\mathbf{y}_t$ being generated by the speaker model λ consisting of L mixtures is

$$P(\mathbf{y}_t|\lambda) = \sum_{l=1}^{L} P(l|\lambda) \prod_{b\in rel} P(y_t(b)|l,\lambda) \prod_{b\in unrel} P(y_t(b)|l,\lambda) \quad (3)$$

where $P(l)$ is the weight of the l^{th} mixture component, and $P(y_t(b)|l)$ is the probability of the b^{th} element of the feature vector given mixture l. The marginalization-based MFT model eliminates the contribution of the unreliable features from the overall probability by integrating them out, hence the product over the unreliable features in Eq. 3 equals to one and the overall probability is then calculated as

$$P(\mathbf{y}_t|\lambda) = \sum_{l=1}^{L} P(l|\lambda) \prod_{b\in rel} P(y_t(b)|l,\lambda). \quad (4)$$

In order to apply the MFT marginalization model, the noise-corruption needs to be localised into several features. This makes the full-band cepstral coefficients, i.e., applying DCT over the entire vector of log filter-bank energies (logFBEs), unsuitable parameterization. The logFBEs may be used, however, they suffer from a high correlation between the features, which makes the diagonal covariance matrix modelling not appropriate. The parameterizations often used in the MFT model are the sub-band cepstral coefficients and frequency-filtered logFBEs (FF-logFBEs), e.g., [13] [14]. 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 have been shown to obtain similar performance as the standard full-band cepstral coefficients [15], while having the advantage of retaining the noise-corruption localized.

3.2 Mask estimation for filter-bank channels

The MFT model described in the previous section requires a mask, each element of which indicates whether the corresponding feature is reliable or unreliable.

The masks for the static features which were evaluated in this paper are presented in the following sections. It is considered that the X, Y and N denote an FBE of the clean speech, noisy speech and noise, respectively.

We analysed three ways of dealing with the delta features: using all delta features, *strict* and *majority* delta masks. Based on the *strict* and *majority* delta-mask a delta feature is defined as reliable only if all and majority of the static features used for computing the delta feature are reliable, respectively.

3.2.1 Oracle mask

Oracle mask is derived based on full a-priori knowledge of the noise and clean speech signal. As such, oracle mask indicates an upper bound performance, and signifies the quality of a mask obtained by an estimation method. It is constructed, by comparing the FBEs of clean speech and noise, as

$$m_{Oracle}(t,b) = 1 \quad \text{if} \quad 10\log\big(X(t,b)/N(t,b)\big) > \gamma. \quad (5)$$

Threshold γ was set to 0dB, which corresponds to the mask value being 1 when a filter-bank (FB) channel is dominated by the speech signal rather than noise – it has been demonstrated that such mask is related to the human auditory masking phenomenon [16].

3.2.2 Voicing masks

It has been demonstrated in [12] that the voicing distance vd^{fb} is related to the local SNR of a voiced FB-channel corrupted by white noise. Based on this, the voicing distance can be used to define the *voiced-feature mask* as

$$m_{VoicedFeat}(t,b) = 1 \quad \text{if} \quad vd^{fb}(t,b) < \beta \qquad (6)$$

where the threshold β was set to 0.21 based on the analysis presented in [12].

In order to evaluate the quality of the voiced-feature masks estimated by the proposed method, i.e., the effect of errors in the voicing information estimation on the speaker identification performance, we defined *voiced-oracle mask* for an FB-channel as 1 if and only if the channel is estimated as voiced on clean data and its oracle mask is 1 on noisy data.

In order to analyse the significance of estimating the voicing information of each FB-channel, we also performed experiments using the voicing information only about an entire frame. To do this, we defined a *voiced-frame mask* as 1 for all features when the frame is voiced; a frame was assigned as voiced if there are at least three FB-channels detected as voiced by the proposed method (as this gave best results in overall).

3.2.3 Noise-estimate-based masks

Noise-estimate masks were obtained based on the estimate of the mean vector of FBEs of noise, denoted as μ_N. Two methods for estimation of μ_N were used. In the first method (*noiseEst1*) the μ_N was estimated based on the first ten frames of each utterance which do not contain speech signal. In the second method (*noiseEst2*) the noise mean for each FB-channel was adapted at each frame of the utterance using a similar procedure as proposed in [17]; the adaptation was performed only when $Y(t,b) < \beta\mu_N(t-1,b)$ by

$$\mu_N(t,b) = \alpha\mu_N(t-1,b) + (1-\alpha)Y(t,b) \qquad (7)$$

where the α and β was set to 0.9 and 2, respectively. In both noise-estimate masks, an FB-channel was then assigned a mask value based on

$$m_{NoiseEst}(t,b) = 1 \quad \text{if} \quad 10\log\big(Y(t,b)/\mu_N(b)\big) > \gamma \quad (8)$$

where the threshold γ was set to 3dB.

4. EXPERIMENTS AND RESULTS

The experimental evaluation of the above-mentioned systems was performed for a speaker identification task.

4.1 Experimental set-up

Experiments were performed on the TIMIT database, down sampled to 8kHz. 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 Aurora2 database, at global SNRs equal to 20dB, 15dB, 10dB and 5dB, respectively.

The frequency-filtered logarithm filter-bank energies [15] were used as speech feature representation, due to their suitability for missing-feature based recognition. These were obtained with the following parameter set-up: frames of 32 ms length with an overlap of 10 ms between frames were used; both pre-emphasis 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 $H(z)=z-z^{-1}$ [15]. A feature vector consisting of 18 elements was obtained (the edge values were excluded). A frequency-filtered (FF) feature was assigned as reliable only if both of the filter-bank channels involved in the calculation of the FF-feature were reliable, and unreliable otherwise. In order to include dynamic spectral information, the first-order delta parameters were added to the static FF-feature vector.

The speaker identification system was based on Gaussian mixture model (GMM) with 32 mixture-components for each speaker, which was constructed using the HTK software [18]. 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.

4.2 Experimental results

4.2.1 Analysis of the delta masks

Here we analyse the effect of marginalizing the delta features on the speaker identification performance. This was performed because it has been observed in speech recognition task that due to paucity of the reliable static features the use of the delta mask (and its type) may be task dependent, e.g., using all deltas may perform better than strict delta mask. Experiments were performed with the MFT model that marginalizes only static features and uses all the delta features (denoted as DeltaAll), and models that use the majority and strict masks for the delta features (denoted as DeltaMajor and DeltaStrict, respectively). This analysis is presented here based on using the oracle mask for the static features, defined in Section 3.2.1. The results of experiments are presented for clean and noisy speech in Table 1. It can be seen that, except for Subway noise at 20dB SNR, the both ways of marginalization of the delta features in addition to

the static features give substantially better recognition accuracy for all noisy conditions. Both the delta-strict and delta-majority masks produce similar results. The performance of the baseline model was included for comparison and this can be seen as performing poorly. Note that similar effect of delta masks presented here with the oracle static mask were also observed in tests with using other static masks (as used in the following sections). Based on these results the delta-strict mask is used in all the following experiments.

Table 1: *Speaker identification accuracy obtained by using the baseline model and MFT-model employing the oracle static mask and various delta masks.*

Speech type	SNR [dB]	Baseline	MFT – Delta Mask		
			All	Majority	Strict
White	20	56.0	75.5	81.0	80.0
	15	33.5	53.5	66.5	65.0
	10	20.0	40.5	58.5	59.5
	5	7.5	18.5	43.5	42.5
Subway	20	58.0	81.0	83.0	80.0
	15	36.0	64.0	73.5	74.0
	10	20.0	43.5	58.5	61.5
	5	5.5	25.5	47.0	48.5

4.2.2 Evaluation of the voicing masks

This section analyses the speaker identification performance when the voicing information obtained by the proposed method is employed in the MFT model.

First we analyse the effect of using only a frame-level voicing information. Experimental results are presented in Table 2 under the column 'Frame-level'. It can be seen that for clean speech marginalizing features corresponding to frames detected as unvoiced causes only slight decrease in the performance in comparison to the baseline model using all frames. For noisy speech, slight positive effect on the performance is observed in the case of White noise, however, significant improvement is achieved in the case of Subway noise.

Table 2: *Speaker identification accuracy obtained by using the baseline model and MFT-model employing various voicing masks: frame-level mask 'voiced-frame' and feature-level masks 'voiced-feature' and 'voiced-oracle'.*

Speech type	SNR [dB]	Baseline	Voiced Mask		
			Frame level	Feature-level	
				Estim	Oracle
Clean	-	96.0	93.5	83.5	83.5
White	20	56.0	60.5	68.0	71.0
	15	33.5	38.5	59.5	62.0
	10	20.0	20.5	51.5	58.0
	5	7.5	13.0	36.0	40.0
Subway	20	58.0	72.5	71.5	72.5
	15	36.0	56.5	64.0	66.5
	10	20.0	33.5	55.5	57.0
	5	5.5	18.5	44.0	47.5

Now, we explore the effect of having the voicing information for individual filter-bank channels. The results of experiments using the feature-level masks, estimated 'voiced-feature' and 'voiced-oracle', are presented in the last two columns of Table 2. First, the performance obtained with the estimated feature-level voicing mask is compared to the one with frame-level voicing mask. It can be seen that in clean speech the performance drops considerably when using the feature-level voicing mask, which is due to marginalizing out the reliable unvoiced features in the voiced frames. On the other hand, we can see that incorporating the voicing information of individual filter-bank channels gives substantial improvements in the identification accuracy in the case of noisy speech (except the 20dB Subway).

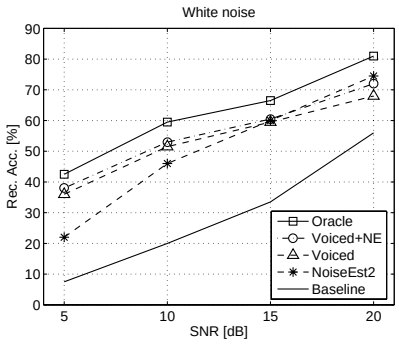
Finally, the performance of the estimated feature-level voicing mask is compared to the one with voiced-oracle mask (defined in Section 3.2.2). It can be observed that for all noisy conditions the recognition accuracy achieved by using the estimated mask is, indeed, very similar to using the voiced-oracle mask which utilizes full a-priori knowledge about the noise.

4.2.3 Evaluation of the voicing mask combined with noise-estimate mask

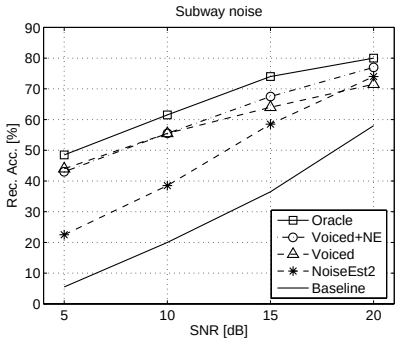
Comparing the oracle results in Table 1 with the voiced-oracle results in Table 2, it can be seen that an inclusion of the reliable unvoiced features can provide a considerable performance improvement (especially at high SNRs). In this section we attempt to utilize the reliable features among the unvoiced features. This could be performed, for instance, by exploiting some characteristic properties of unvoiced speech signal or based on using noise-estimate. We employed here the noise-estimate mask in combination with the voicing mask. Such combined mask could be obtained by a soft weighting of the contribution of each mask which would then be considered in a modified probability calculation of the MFT model. A combined mask is formed, for simplicity, by using all voiced features detected by the voicing mask and the unvoiced features of the voiced frames detected as reliable by the noise-estimate mask (noiseEst2). Experimental results with the combined mask, compared also to other masks, for noisy speech are presented in Figure 2. Note that for clarity only the results with noiseEst2 mask are presented as both masks noiseEst1 and noiseEst2 performed similarly in the case of White noise, however, the noiseEst2 mask gave better results (around 5% on average) in Subway noise. It can be seen that using the combined mask gives similar results to the noiseEst2 mask at high SNRs in the case of White noise, however, performs significantly better at low SNRs in both White and especially Subway noise. Using the combined mask in comparison to the voiced-feature mask gives similar performance at low SNRs while provides improvement in the high SNRs speech. The results for clean speech (not presented in the figure) showed an improvement in the identification accuracy from 83.5% when using only the voiced-feature mask (see Table 2) to 92.5% when using the combined mask. In future work, we plan to incorporate a more complex method for detection of the reliable unvoiced features.

5. CONCLUSION

In this paper, we presented a novel method for estimation of the voicing information of frequency-regions of speech spectra and analysis of employment of the voicing information, in the form of a reliability mask, into a missing-



(a)



(b)

Figure 2: *Speaker identification accuracy obtained by the MFT-model employing various masks for speech corrupted by White (a) and Subway (b) noise at various SNRs.*

feature based speaker identification system. The experimental evaluation was performed using the TIMIT database on clean speech and speech corrupted by a stationary and real-world non-stationary noises at various SNRs. The employment of voicing mask estimated by the proposed method showed identification performance very close to using the oracle voicing mask obtained based on the full a-priori knowledge of noise. A combination of the voicing mask with a noise-estimate mask was explored in order to obtain an estimate of reliable unvoiced features and this showed further improvement in the identification accuracy. The obtained results are significantly higher than using the baseline model and, in many cases, noise-estimate-based masks and indeed considerably close to the upper bound results defined by using the oracle SNR mask. The performance can still be further improved by employing a more accurate estimation of reliability of unvoiced features, and modifying the probability calculation, which is our future work.

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

REFERENCES

[1] S.F. Boll, "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.

[2] S.V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, John Wiley & Sons, 2005.

[3] X. Zou, P. Jančovič, J. Liu, and M. Köküer, "ICA-based MAP Algorithm for Speech Signal Enhancement," *ICASSP, Honolulu, Hawaii*, p. accepted, 2007.

[4] M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Proc.*, vol. 4, pp. 352–359, 1996.

[5] R.P. Lippmann and B.A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise," *Eurospeech, Rhodes, Greece*, pp. 37–40, 1997.

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

[7] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environment with combined spectral subtraction and missing data theory," *ICASSP, Seattle, WA*, vol. I, pp. 121–124, 1998.

[8] P. Renevey and A. Drygajlo, "Statistical estimation of unreliable features for robust speech recognition," *ICASSP, Istanbul, Turkey*, pp. 1731–1734, 2000.

[9] M.L. Seltzer, B. Raj, and R.M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.

[10] J.P. Eatock and J.S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," *ICASSP, Adelaide, Australia*, vol. I, pp. 133–136, 1994.

[11] X. Zou, P. Jančovič, and J. Liu, "The Effectiveness of ICA-based Representation: Application to Speech Feature Extraction for Noise Robust Speaker Recognition," *European Signal Processing Conference, Florence, Italy*, 2006.

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

[13] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," *ICSLP, Philadelphia, USA*, 1996.

[14] P. Jančovič and J. Ming, "A multi-band approach based on the probabilistic union model and frequency-filtering features for robust speech recognition," *Eurospeech, Aalborg, Denmark*, pp. 1111–1114, 2001.

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

[16] B. C. J. Moore, *An introduction to the psychology of hearing*, Academic Press, San Diego, 5th edition, 2003.

[17] H. Hirsch and C. Erlicher, "Noise estimation techniques for robust speech recognition," *ICASSP, Detroid*, pp. 153–156, May 1995.

[18] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book. V2.2*.