14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

THE EFFECTIVENESS OF ICA-BASED REPRESENTATION: APPLICATION TO SPEECH FEATURE EXTRACTION FOR NOISE ROBUST SPEAKER RECOGNITION

Xin Zou[1], *Peter Jančovič*[1] *and Ju Liu*[2]

[1] Electronic, Electrical & Computer Engineering, University of Birmingham, Birmingham, UK

[2] School of Information Science and Engineering, Shandong University, Shandong, China

{xxz391, p.jancovic}@bham.ac.uk, juliu@sdu.edu.cn

ABSTRACT

In this paper, we present a mathematical derivation demonstrating that feature representation obtained by using the Independent Component Analysis (ICA) is an effective representation for non-Gaussian signals when being both clean and corrupted by Gaussian noise. Our findings are experimentally demonstrated by employing the ICA for speech feature extraction; specifically, the ICA is used to transform the logarithm filter-bank-energies (instead of the DCT which provides MFCC features). The evaluation is presented for a GMM-based speaker identification task on the TIMIT database for clean speech and speech corrupted by white noise. The effectiveness of ICA is analysed individually for signals corresponding to each phoneme. The experimental results show that the ICA-based features can provide significantly better performance than traditional MFCCs and PCA-based features in both clean and noisy speech.

1. INTRODUCTION

In current speech and speaker recognition systems, mel-frequency cepstral coefficients (MFCCs) are the most widely used representation of speech. However, there has been much effort to find a better representation of speech signals.

The Independent Component Analysis (ICA), which has been introduced in the context of blind signal separation, linearly transforms data to be statistically as independent from each other as possible. The ICA has recently been applied also for speech feature extraction. The authors in [1] used ICA in the time-domain to replace the Fourier transform. The authors in [2] showed that ICA applied in the log spectral domain provides features similar to cepstral coefficients. In [3] [4] the ICA was used on the log filter-bank-energy feature vector for speech recognition. In the above research, there was little attention devoted to the motivation for the use of ICA for feature extraction – indeed, this has usually been addressed only by stating the ability of the ICA to exploit not only the second-order but higher-order statistics. Besides, the experimental evaluations have been demonstrated only on clean speech with usually only little improvement over the traditional MFCCs.

In this paper, we present a mathematical derivation that shows that ICA-based independent features are an effective representation of non-Gaussian signals when being both clean and corrupted by Gaussian white noise. The derivation is based on calculating a mismatch between the density distribution of the original clean data and observed data represented by our model. In order to experimentally demonstrate our findings, we employed the ICA in the log mel-scaled spectral domain for speech feature extraction, i.e. ICA replaces the traditionally used DCT in the MFCCs calculation. The experiments are performed for a speaker identification task on the TIMIT database using the Gaussian Mixture model (GMM). The performance is analysed for clean speech and speech corrupted by white noise. The experiments are first performed with using all signal-frames and then the effectiveness of ICA on signals corresponding to individual phonemes is explored. The experimental results show that the ICA-based features provide significantly better performance than the traditional MFCCs and PCA-based features for both clean and noisy speech when using signal-frames corresponding to vowel and semi-vowel categories. For instance, when using signal-frames corresponding to a subset of vowels, the best speaker-identification accuracy for speech corrupted by white noise at 10dB obtained by the ICA-based features was 50%, while the traditional MFCCs gave only 30% accuracy.

2. THE EFFECTIVENESS OF ICA-BASED REPRESENTATION

In this section we demonstrate that the ICA-based features are an effective representation for non-Gaussian signals.

In standard signal representation, the input signal is processed within a linear framework, which can be modelled in terms of a linear superposition of basis functions ϕ_i mixed with weights a_i which will be used as features. If we assume some additive Gaussian white noise $\mathbf{v}$, the signal model will be:

$$\mathbf{x} = \sum_i a_i \phi_i + \mathbf{v} \tag{1}$$

The probability density function of the signal $\mathbf{x}$ from a particular choice of $\mathbf{a}$ is given by $p(\mathbf{x}|\mathbf{\Phi},\mathbf{a}) = p_{\mathbf{v}}(\mathbf{x}-\mathbf{\Phi a})$, where $p_{\mathbf{v}}(\mathbf{x}-\mathbf{\Phi a})$ is the density of noise evaluated at $\mathbf{x}-\mathbf{\Phi a}$. Note that for clean signal, the noise can be treated as Gaussian distribution with zero variance, hence the distribution function would be a delta function.

The effectiveness of the representation can be analysed by assessing how well the density distribution of noisy signal represented by the model $p(\mathbf{x}|\mathbf{\Phi})$ matches the density distribution $p^*(\mathbf{x})$ of the original clean signal. As a measure of the mismatch, we apply Bhattacharyya distance (BD) between the two distributions (the distance is zero when the two distributions are identical):

$$BD = -\ln \int \sqrt{p^*(\mathbf{x}) p(\mathbf{x}|\mathbf{\Phi})}\, d\mathbf{x} = -\ln \langle \sqrt{\frac{p(\mathbf{x}|\mathbf{\Phi})}{p^*(\mathbf{x})}} \rangle \tag{2}$$

where $\langle\rangle$ denotes expectation. By Jensen's inequality $f(E(y)) \leq E(f(y))$, and because $p^*(\mathbf{x})$ is fixed, so minimization of the distance equals to the maximization of expectation $\langle \ln p(\mathbf{x}|\Phi)\rangle$:

$$\min(BD) \Leftrightarrow \min_{\Phi} \langle -\frac{1}{2} \ln \frac{p(\mathbf{x}|\Phi)}{p^*(\mathbf{x})} \rangle \Leftrightarrow \max_{\Phi} \langle \ln p(\mathbf{x}|\Phi)\rangle \quad (3)$$

where $p(\mathbf{x}|\Phi) = \int p(\mathbf{x}|\Phi,\mathbf{a})p(\mathbf{a})$. Unfortunately, the evaluation of $p(\mathbf{x}|\Phi)$ requires integrating over all possible values of $\mathbf{a}$, which is in general intractable. In order to simplify the calculation, assuming that the function inside the integral (i.e. $p(\mathbf{x}|\Phi,\mathbf{a})p(\mathbf{a})$) has a maximum, then the integral may be approximated by evaluating the function at its maximum [5]. Therefore, our goal then becomes:

$$\begin{aligned} \min(BD) &\Leftrightarrow \max_{\Phi} \langle \ln \int p(\mathbf{x}|\Phi,\mathbf{a})p(\mathbf{a})\rangle \\ &\Leftrightarrow \max_{\Phi}(\max_{\mathbf{a}} \langle \ln (p(\mathbf{x}|\Phi,\mathbf{a})p(\mathbf{a}))\rangle) \end{aligned} \quad (4)$$

where the term $\langle \ln(p(\mathbf{x}|\Phi,\mathbf{a})p(\mathbf{a}))\rangle = \langle \ln p(\mathbf{x}|\Phi,\mathbf{a})\rangle + \langle \ln p(\mathbf{a})\rangle$.

Let us first consider the term $\langle \ln p(\mathbf{a})\rangle$. According to [6], the maximization problem can be solved within the blind framework. One possible density function approximation of coefficients a_i can be constructed in the following way:

$$\widetilde{p}^{+}(a_i) = \exp(\alpha_1 - \log(\cosh a_i)) \quad (5)$$

$$\widetilde{p}^{-}(a_i) = \exp(\alpha_2 + \log(\cosh a_i) - a_i^2/2) \quad (6)$$

where α_1, α_2 are constants and $\widetilde{p}^{+}$ and $\widetilde{p}^{-}$ is a supergaussian and subgaussian density, respectively. Let us denote $E\{G(\mathbf{a})\} = \langle G(\mathbf{a})\rangle = \langle \ln p(\mathbf{a})\rangle$. In basic ICA model, the original observations can be expressed by the linear function of some independent components, i.e. $\mathbf{x}^* = \mathbf{M}\mathbf{s}$, where $\mathbf{s}$ are the independent components which are assumed to be fixed in our case, $\mathbf{M}$ is a mixing matrix. By our model $\mathbf{x}^* = \Phi\mathbf{a}$, so we have $\mathbf{a} = \Phi^{-1}\mathbf{M}\mathbf{s}$. Let us denote $\Phi^{-1}\mathbf{M}$ by $\mathbf{q}$, and $E\{G(\mathbf{a})\}$ by $H(\mathbf{q})$. Without loss of generality, it is enough to analyze the function's stability of the point $\mathbf{q} = \mathbf{e}_1$, where $\mathbf{e}_1 = (1,0,\cdots,0)$. According to [6], the Taylor extension by making a small perturbation $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots)$ at point $\mathbf{e}_1$ can be approximated in the form of:

$$H(\mathbf{e}_1 + \varepsilon) \approx H(\mathbf{e}_1) + \frac{1}{2} E\{g'(s_1) - s_1 g(s_1)\} \sum_{i>1} \varepsilon_i^2 \quad (7)$$

where g and g' is the first and second derivative of G. The use of Eq.5 and Eq.6 guarantee that $E\{g'(s_1) - s_1 g(s_1)\} < 0$. Therefore, Eq.7 implies that the maximization of $\langle \ln p(\mathbf{a})\rangle$ will finally converge at the point where the components of $\mathbf{a}$ are independent.

Now, we consider the term $\langle \ln p(\mathbf{x}|\Phi,\mathbf{a})\rangle$. Similar procedure as above leads to the approximation of $\langle \ln p(\mathbf{x}|\Phi,\mathbf{a})\rangle$ in the form of Eq.7. As shown earlier, $p(\mathbf{x}|\Phi,\mathbf{a}) = p_{\mathbf{v}}(\mathbf{x} - \Phi\mathbf{a})$, where $p_{\mathbf{v}}$ is the density of noise. Considering a Gaussian noise, it can easily be shown that the term $E\{g'(s_1) - s_1 g(s_1)\} = 0$. So, there is no extremum at all for the term $\langle \ln p(\mathbf{x}|\Phi,\mathbf{a})\rangle$.

Based on the above, we can draw the conclusion that for our model, the function $\langle \ln(p(\mathbf{x}|\Phi,\mathbf{a})p(\mathbf{a}))\rangle$ has a maximum at the point where the coefficients $\mathbf{a}$ of the basis functions Φ are independent. As such, the minimum mismatch between the distribution of clean non-Gaussian signal and corrupted by a Gaussian noise is achieved by independent representation $\mathbf{a}$. The basis functions Φ, whose coefficients are independent from each other, can be obtained by optimization of a criteria, such as, kurtosis [7], negentropy [6] or mutual information [8]; these are referred to as ICA algorithms.

3. SPEECH FEATURE EXTRACTION: ICA-BASED TRANSFORMATION OF LOG FILTER-BANK-ENERGIES

The theoretical results presented in the previous section could be demonstrated by employment of the ICA in various stages of speech feature extraction. This section briefly describes the employment we adopted in this paper.

A block diagram of a typical frame-based speech feature extraction is depicted in Figure 1. This consists of dividing the speech signal into frames, computing the short-term magnitude spectra and estimating the envelope of the spectra by using mel-scaled filter-bank analysis. This is usually compressed by logarithm function, giving a vector of logarithm filter-bank-energies (logFBEs). Then, a linear transformation is usually applied in order to decorrelate the vector of logFBEs.

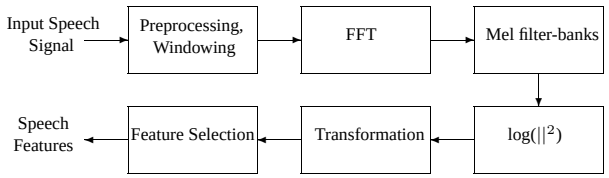

Figure 1: *Block diagram of a typical frame-based speech feature extraction process.*

3.1 ICA-based transformation of log filter-bank-energies

Traditionally, the discrete cosine transform (DCT) is applied in the transformation step in the Figure 1, producing mel-frequency cepstral coefficients (MFCCs). Motivated by the theoretical results presented in Section 2, we explore the use of ICA as a replacement of DCT. To extract independent feature vectors, ICA algorithm is applied to a set of noise-free training logFBE vectors $\mathbf{x}$ to obtain independent features $\mathbf{a}$ and the unmixing matrix $\mathbf{W}$. The matrix $\mathbf{W}$ is the inverse of mixing matrix Φ, i.e. the columns of Φ are the ICA basis functions. The trained unmixing matrix $\mathbf{W}$ is applied on each frame logFBE feature vector $\mathbf{x}$ of the testing data to yield the features used for recognition, i.e.,

$$\mathbf{a} = \mathbf{W} \cdot \mathbf{x} \quad (8)$$

Since the input signal $\mathbf{x}$ is usually pre-whitened by PCA algorithm prior to performing the ICA algorithm, the matrix $\mathbf{W}$ in Eq.8 consists of two matrices (i.e. PCA and ICA matrix).

Note that Gaussian noise added to a signal-frame in the time-domain will not be equivalent to Gaussian in the logFBE vector. However, for a given filter-bank channel the logFBEs over frame-time are, based on the central limit theorem, distributed close to Gaussian.

3.2 Feature Selection

The feature representation of signal may be based on using only a subset of features, i.e. the dimension of the transformed feature vector $\mathbf{a}$ (obtained by Eq.8) may be lower than original vector $\mathbf{x}$. Two feature selection methods were used.

3.2.1 Based on the variance described by PCA basis

Typically, the observed data are whitened by PCA before the application of ICA training algorithm. As such, the feature selection can be performed based on the amount of variance reflected by the eigenvalues, which is a standard method used in data analysis.

3.2.2 Based on the L_2-norm of ICA basis

The feature selection can also be based on the L_2-norm of the ICA basis vectors. The L_2-norm of a basis vector expresses the contribution of the basis to the speech signals and by selecting those basis vectors with high L_2-norm, the reconstruction mean squared error will be minimized. Figure 2 shows the ICA basis vectors ordered based on the L_2-norm. The value of the norm of each basis is depicted on Figure 3. It can be seen in Figure 2 that, unlike the DCT basis, many of the ICA basis are localized.

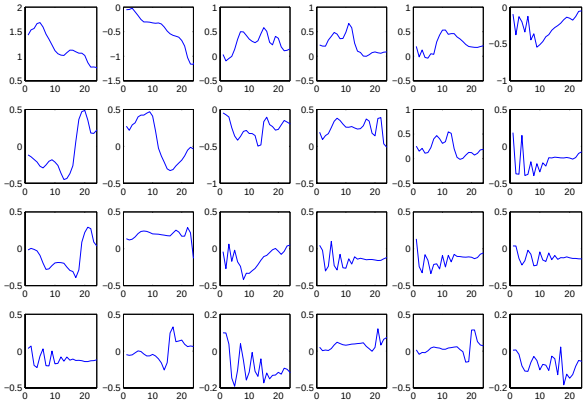

Figure 2: *ICA basis functions ordered based on the L_2 norm (by rows from left to right).*

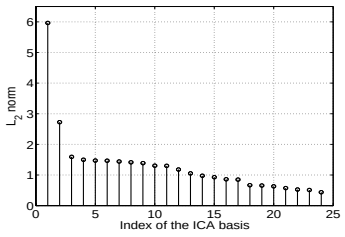

Figure 3: *The L_2 norm of the ICA basis functions.*

4. EXPERIMENTS AND RESULTS

The experimental evaluation of the ICA-based features was performed for a speaker identification task.

4.1 Experimental set-up

The experiments were performed on the TIMIT database, downsampled to 8kHz. The 100 speakers (consisting of 64 male and 36 female) from the test subset were selected. The ICA transformation matrix was estimated on the clean speech by using the JADE algorithm [7]. The training set consisted of eight sentences ('si' and 'sx') for each speaker and testing was performed using two sentences ('sa').

The speech signal was divided into frames of 30 ms with an overlap of 10 ms between frames. Both preemphasis and Hamming window were applied to each frame. For each frame, Mel-scaled filter bank analysis with 24 channels was performed. These were transformed by using the DCT to obtain the traditional MFCC features and by using the PCA-based and ICA-based transformation. In each case, the final feature vector consisted of 18 components with first-order deltas resulting in a 36-component feature vector. The speaker recognition system is based on 256 mixtures Gaussian mixture modelling (GMM), which was constructed using the HTK software [9]. 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. For recognition, the testing set was corrupted by Gaussian white noise at global SNR equal to 20dB and 10dB, respectively.

4.2 Experimental results

First experiments were performed by using features obtained from all the frames of the signal. The experimental results for clean speech and speech corrupted by white noise are presented in Table 1. The results obtained with the ICA-based features selected based on the norm of the ICA basis are presented (these results differed by a maximum of 0.5% to those selected based on the PCA eigenvalues). It can be seen that in the case of clean speech, the ICA- and PCA-based features performed similarly; they both obtained better recognition accuracy than using the MFCCs. In the case of noisy speech, the ICA-based features well outperformed both the MFCCs and PCA features.

Table 1: *Speaker-identification accuracy obtained by using MFCC, PCA- and ICA-based features for clean speech and speech corrupted by white noise at various SNRs when using all signal-frames.*

Speech type SNR [dB]	Features		
	MFCC	PCA	ICA
clean	95.5	99.0	98.0
20	57.5	59.5	63.0
10	11.0	8.5	16.0

Next experiments were performed to explore the effectiveness of ICA on signals corresponding to individual phonemes, i.e. the speaker-identification is based on using only the signal-frames corresponding to each individual phoneme. The results are presented in Table 2 (note that some phonemes occurring rarely in the data are not included). Let us first analyse the results for clean speech. As can be seen from Table 2, all the feature representations provided similar performance when using stops and fricatives. On the other side, the ICA features gave significant improvement (both over the MFCCs and PCA) for phonemes that

Table 2: *Speaker-identification accuracy obtained by using signal-frames corresponding to each individual phoneme. Clean speech and speech corrupted by white noise at SNR=10dB.*

Phoneme			Clean Speech			Noisy Speech		
category	label	example	MFCC	PCA	ICA	MFCC	PCA	ICA
Stops	t	top	18	22	15	1.7	1.7	5
	dx	butter	18.4	22	23	3	4	8
	k	kick	14	15	16	3	2	1
	g	gag	8	12	14	4	1	2
	gcl	(g closure)	43	43	42	3	1	3
	kcl	(k closure)	40	48	51	1	0	4
Nasals	m	mom	24	28	33	4	2	4
	n	non	60.6	69	65	1	5	7
Fricatives	dh	they	15	16	7	1	3	1
	s	sis	42	57	45	4	2	3
	sh	shoe	33	30	17	3	1	2
Semivowels,	l	led	32	46	46	8	6	12
Glides	r	red	49	39	51	5	8	14
	y	yet	24	26	20	2	4	5
	hv	ahead	7.5	8.6	9.7	5.4	4.3	3
Vowels	iy	beat	67	74	78	16	11	19
	ih	bit	30	43	48	4	7.4	9.6
	eh	bet	21.7	29	36	9.6	16	16
	ae	bat	52	63	66	24	22	34
	ix	roses	51	54	60	6	11	18
	ux	toot	17.4	23	34	5.4	6.5	6.5
	ao	about	35	39	40	8	9	14
	aa	cot	22.4	41	34	14	18	18
	ay	bite	27.6	31	31	9	12	9.2
	oy	boy	23.8	29	38	15	8.8	15
	ow	boat	32.7	36	37	15	21	19

belong to the vowel, nasal and semi-vowel categories. Now, let us look at results for noisy speech. It can be seen that, regardless the feature representation, all phonemes but vowels (and some semivowels) provide very poor recognition performance. This is probably because the signals for these phonemes are typically of much lower energy than vowels, and as such they are more affected by the noise. For the vowel category, the results in Table 2 show that the ICA-based features provide significantly higher recognition performance than both MFCCs and PCA-based features. These results confirm our theoretical derivations.

Motivated by the results of Table 2, we performed experiments when using signal-frames corresponding to a subset of specified phonemes. These experiments show the speaker identification accuracy that could be obtained when the used phonemes are correctly identified. The following set of 18 phonemes, ordered based on the performance they provided on noisy speech as presented in Table 2, were chosen: {'ae', 'iy', 'ow', 'aa', 'ix', 'eh', 'oy', 'ao', 'r', 'l', 'ih' , 'ay', 'n', 'ux', 'axr', 'y', 'w', 'm'}. The experiments were performed by using signal-frames corresponding to phoneme subsets from the above list which includes first *n* phonemes, i.e. phoneme subset one corresponds to using the signal-frames of 'ae' only, subset two includes 'ae' and 'iy', etc. The experimental results for speech corrupted by white noise at SNR=10dB are shown on Figure 4, in which the x-axis indicates the phoneme subset used. It can be seen that the ICA-based features significantly outperform both the MFCCs and PCA-based features. The best recognition performance obtained by ICA was 50%, while the PCA and MFCCs obtained only 27% and 30%, respectively.

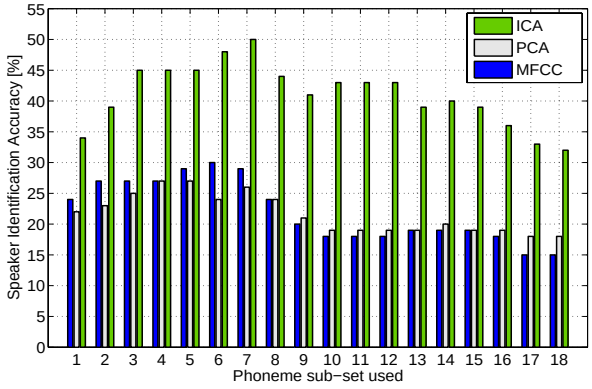

Figure 4: *Speaker-identification accuracy for speech corrupted by white noise at SNR=10dB when using signal-frames of specified sub-set of phonemes.*

5. CONCLUSION

In this paper, we presented a mathematical derivation that demonstrated that the Independent Component Analysis (ICA) provides an effective feature representation for non-Gaussian signals when being both clean and corrupted

by Gaussian noise. The derivation is based on calculation of a mismatch between the density distribution of the original clean and noise-corrupted signal represented by our model. Theoretical findings were experimentally demonstrated by employment of the ICA for speech feature extraction – specifically, we used the ICA to transform the logarithm filter-bank-energies (i.e. the ICA replaced DCT used in calculation of the standard MFCCs). The obtained ICA-based features were employed in a GMM-based speaker identification system. The experimental evaluations were performed on the TIMIT database on clean signal and signal corrupted by white noise. We explored the effectiveness of the ICA on signals corresponding to individual phonemes. The experimental results showed significant improvement by using the ICA-based features in comparison to both the MFCCs and PCA features for both clean speech and speech corrupted by white noise. The best identification accuracy for speech corrupted by white noise at 10dB (by using signal-frames corresponding to a selected subset of phonemes) was 50% when using the ICA-based features, while only 30% and 27% when using the MFCCs and PCA-based features, respectively. In our future work, we will investigate the effect of ICA-based feature representation for signals corrupted by a non-Gaussian noise.

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

REFERENCES

[1] J.H. Lee, T.Y. Lee, T.W. Jung, and S.Y. Lee, "Speech feature extraction using independent component analysis," *ICASSP, Istanbul, Turkey*, pp. 1631–1634, 2000.

[2] J. Rosca and A. Kofmehl, "Cepstrum-like ica representation for text independent speaker recognition," *Int. Symposium on Independent Component Analysis and Blind Signal Separation, Nara, Japan*, pp. 999–1004, 2003.

[3] O.-W. Kwon and T.-W. Lee, "Phoneme recognition using ica-based feature extraction and transformation," *Signal Processing*, vol. 84, pp. 1005–1019, 2004.

[4] L. Potamitis, N. Fakotakis, and G. Kokkinakis, "Independent component analysis applied to feature extraction for robust automatic speech recognition," *IEE Electronic Letters*, vol. 36, no. 23, pp. 1977–1978, Nov. 2000.

[5] Bruno A.Olshausen and David J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1," *Vision Research*, vol. 37, pp. 3311–3325, 1998.

[6] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley & Sons, Inc., 2001.

[7] J.F. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.

[8] A.J. Bell and T.J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

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