VOICING-CHARACTER ESTIMATION OF SPEECH SPECTRA: APPLICATION TO NOISE ROBUST SPEECH RECOGNITION

Peter Jančovič[1] *and Münevver Köküer*[2]

[1] Electronic, Electrical & Computer Engineering, University of Birmingham, B15 2TT, Birmingham, UK;

[2] School of Mathematical and Information Sciences, Coventry University, Coventry, UK

p.jancovic@bham.ac.uk, m.kokuer@coventry.ac.uk

ABSTRACT

This paper presents a novel method for estimating the voicing-character of speech spectra, demonstrates its employment in noise robust ASR and proposes a modified calculation of filter-bank energies. The proposed voicing-character estimation is based on calculation of a similarity between the shape of the signal short-term magnitude spectra around spectral peaks and spectra of the frame-analysis window. The similarity is weighted by the signal magnitude spectra to reflect the filter-bank analysis typically used in feature extraction for speech recognition. The experimental results show less than 5% false-acceptance and false-rejection errors in detection of voiced filter-bank channels in speech signal corrupted by white noise at 10dB local SNR. The recognition results obtained by a missing-feature based ASR system using features estimated as voiced by the proposed method are very similar to using oracle voicing-label obtained by full a-priori knowledge of noise. The employment of features obtained by a modified calculation of filter-bank energies shows further improvements in the recognition accuracy.

1. INTRODUCTION

The short-term spectra of speech signal reflects information about both the vocal-tract filter and source-signal used to produce the speech. The spectra of voiced speech sounds are characterized by the presence of a harmonic structure, while unvoiced speech sounds have a stochastic spectral character. Thus, the estimation of whether a region of short-term spectra has a harmonic or stochastic character, which we refer to as voicing-character information, can provide important information that could be exploited for the development of more appropriate speech pattern processing techniques.

The information about the voicing-character of speech spectra has been mainly exploited in speech coding and speech synthesis research. In [1] the voicing-character is estimated based on the closeness of fit between the original and synthetic spectrum at each harmonic of the estimated fundamental frequency (F0). The authors in [2] estimate the voicing-character of a 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 voicing-character of speech spectra can also be estimated by decomposing speech signal into harmonic and stochastic components, by using some decomposition methods (e.g. [3], [4]), and then evaluating the ratio of energy of the harmonic and stochastic components. The voicing-character estimation was not the primary aim of the above methods and as such performance evaluation of this was not provided. Besides, they require estimation of the F0, which may be difficult to estimate accurately in noisy speech.

In this paper, we present a method for voicing-character estimation of speech spectra that is particularly applicable to speech pattern processing. In the first step, a similarity, which we refer to as voicing-distance, between the shape of signal short-term spectra around each spectral peak and spectra of the frame-analysis window is calculated. Then, a voicing-distance associated with a filter-bank channel is computed as an average of voicing-distances weighted by corresponding spectral magnitude values. This reflects the filter-bank analysis typically used in feature extraction for speech pattern processing. The experimental evaluation of the proposed method is presented in terms of false-rejection and false-acceptance errors in a voiced/non-voiced character estimation of filter-bank channels. Then, the performance is evaluated in terms of recognition accuracy when the voicing-character information is employed in a missing-feature based ASR system [5]. Finally, we present a modification to the calculation of filter-bank energies. The results indicate that the proposed methods can provide a significant improvement in recognition performance, while requiring no information about the noise.

2. VOICING-CHARACTER ESTIMATION OF SPEECH SPECTRA BASED ON SPECTRAL SHAPE

2.1. Motivation

Based on the source-filter model [6] of speech production, the source signal used for production of voiced speech is quasi-periodic. Thus, the short-term Fourier spectrum of voiced speech segment can be represented as summation of scaled and shifted versions of the Fourier transform of frame-window function $W(\omega)$, i.e.

$$S(\omega) = \sum_{h=0}^{H} A_h \cdot W\big(\omega - (h+1)\cdot\omega_0\big) \qquad (1)$$

where ω_0 is the fundamental frequency, and $A_h = |A_h| \cdot e^{j\phi_h}$ represents the complex amplitude (the ϕ_h being phase) of the h^{th} harmonic component (i.e. sine-wave). In Eq. 1, for a given harmonic frequency, the contributions of side-lobes of $W(\omega)$ corresponding to other harmonics can be neglected due to their amplitude being much lower than the amplitude of the main-lobe of actual $W(\omega)$. Then, considering that the main-lobes corresponding to adjacent harmonics are well separated (i.e. the fundamental frequency is not extremely low), the shape of the magnitude spectra of voiced speech around each harmonic frequency should follow approximately the shape of the magnitude spectra of the frame analysis window $W(\omega)$.

2.2. Algorithm description

Below are the steps of the proposed algorithm for voicing-character estimation of filter-bank channels:

1) Short-term magnitude-spectra calculation: Samples of a signal frame are weighted by a frame-analysis window function and the FFT is applied on the vector of signal samples expanded by zeros in order to provide a smoother short-term spectral magnitude.

2) Voicing-distance calculation: For each peak of the signal short-term magnitude-spectra, a similarity between the shape of the signal spectra around the peak and the magnitude-spectra of the frame-analysis window is computed – this is referred to as *voicing-distance* and denoted by $vd(k)$. Specifically, we used the Euclidean distance between the logarithm spectra, i.e.

$$vd(k_p) = \left[\frac{1}{2M+1} \sum_{m=-M}^{M} \left(\log_{10} \frac{|S(k_p+m)|}{|W(m)|} \right)^2 \right]^{1/2} \quad (2)$$

where k_p is frequency-index of 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.2. Note that spectral peaks are identified by detection of the changes of the slope of $|S(k)|$ from positive to negative.

3) Voicing-distance calculation for filter-bank channels: The calculation of the voicing-distance for filter-bank channels is carried out in such a way that it reflects the calculation of filter-bank energies typically used to derive features in current speech pattern processing. Hence, the voicing-distance for filter-bank channels is defined as the sum of voicing-distances (associated with frequency components within the region of the filter-bank channel), each being weighted according to the contribution of the frequency component to the overall filter-bank energy, i.e.

$$vd^{fb}(b) = \frac{1}{X(b)} \cdot \sum_{k=k_b}^{k_b+K_b-1} vd(k) \cdot G_b(k) \cdot |S(k)|^2 \quad (3)$$

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 $X(b) = \sum_{k=k_b}^{k_b+K_b-1} G_b(k)|S(k)|^2$, i.e. the overall filter-bank energy value. The Eq. 3 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, or a piece-wise linear interpolation, in which $vd(k) = vd(k_p)$ for $k \in < k_p - M, k_p + M >$ and when these intervals (corresponding to adjacent k_p's) overlap the minimum of voicing-distances is taken, otherwise (i.e. there is a gap) linear interpolation is performed between endings of the intervals. Alternatively, the computation of $vd^{fb}(b)$ can be based on using only the voicing-distances corresponding to peaks. In such a case the summation in Eq. 3 and in calculation of $X(b)$ is only through the $k \in \{k_p, k_b \leq k_p \leq k_b + K_b\}$.

2.2.1. Incorporation of filtering of voicing distances

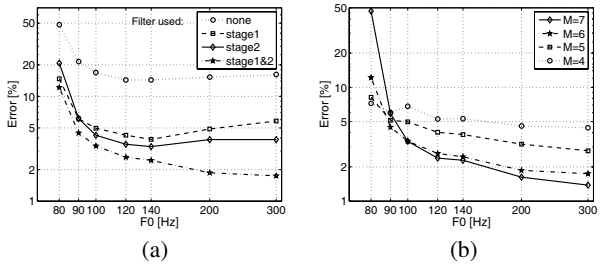
The voicing-distance obtained from Eq. 2 and Eq. 3 may accidentally become of a low value for a non-voiced region or vice versa, i.e. resulting a local outlier. This can be improved by employing a postprocessing (i.e. filtering) of the voicing-distance values. We employed a 2D median filtering due to its effectiveness in eliminating outliers and simplicity. The filtering can be performed on interpolated voicing-distance values $vd(k)$ and/or on voicing-distance values of filter-bank channels $vd^{fb}(b)$. Median filters of size 5×9 and 3×3 were used, respectively.

2.3. Experiments on simulated voiced speech signals

This section discusses and experimentally demonstrates setting for the parameter M and the use of postprocessing on the calculated voicing-distances. The parameter settings were evaluated in terms of an error in voiced/non-voiced character classification of filter-bank channels, i.e. an overlap between distributions of voicing-distance values corresponding to voiced and non-voiced filter-bank channels. Experiments were performed using voiced filter-bank channels corrupted at various local SNRs. The representatives of non-voiced filter-bank channels were obtained by using white noise. The representatives of noise-corrupted voiced filter-bank channels were obtained by adding white noise at various SNRs to simulated voiced speech signals. The simulated voiced speech signals were synthesized as sum of sine-waves (of equal amplitudes) whose frequencies were multiples of the F0, which was set to a value from 80Hz to 300Hz in order to reflect a realistic speech.

Throughout this paper, we consider speech signal sampled at 8kHz. The signal is divided into frames of 256 samples (with an overlap of 80 samples), each frame being obtained by using the Hamming frame-window. The short-term magnitude spectra, obtained by applying the FFT, is passed to Mel-spaced filter-bank analysis with 20 channels. Based on experiments presented in [7], the FFT-size of 1024 points was used in this paper.

An upper-bound for value of the parameter M can be considered to be the half of the main-lobe bandwidth of analysis-window spectra – this corresponds to the value 8 when FFT-size is 1024 points. Choosing a higher value for M, i.e. including components more towards the lower-part of the main-lobe (of the expected shape) in the voicing-distance calculation, may cause the voicing-distance to be easily affected by little noise in the signal or by an overlap of main-lobes corresponding to adjacent harmonics when F0 is low. On the other side, a lower value for M increases the risk of a low voicing-distance being accidentally obtained for a non-voiced spectral peak.



(a) (b)

Fig. 1. *The error in voiced/non-voiced character classification of filter-bank channels when using simulated voiced signals corrupted by white noise as a function of employed filtering of voicing-distances (a), and various values of parameter M (b).*

The voicing-classification error (averaged over local SNRs from 5dB to 20dB) obtained by simulation experiments are presented in Figure 1(a) and (b). First, the effect of filtering of voicing-distances on the error is shown in Figure 1(a) when M is set to 6. It can be

seen that filtering the $vd(k)$ and $vd^{fb}(b)$ values (in the figure, corresponding to stage1 and stage2, respectively) gives similar error, which is significantly lower than using no filter. The error is further decreased when filters at both stages are employed. The effect of setting various values for M on the error rate as a function of voiced signal being produced by various F0 is shown in Figure 1(b) when both filters are employed. Setting the M to 6 or 7 gives considerably high voicing classification error when the F0 is below 90Hz, which is due to the overlap of lower parts of main-lobes of adjacent harmonics. However, these values of M produce much lower error than M set to 4 or 5 when F0 is above 90Hz. As the F0 of speech sounds is in general (and in the speech database used here) rarely below 90Hz, the parameter M was set to 7 for experiments presented in the following section.

2.4. Modified calculation of the filter-bank-energies

Features used for speech pattern processing are usually derived from filter-bank energies (FBEs). The FBEs are typically calculated as

$$X(b) = \sum_k S_w(k) \quad \text{where } S_w(k) = G_b(k)|S(k)|^2 \qquad (4)$$

where the summation is over all $k \in < k_b, k_b + K_b - 1 >$ (see Section 2.2 for description of the notation). For clean speech signal, when the filter-bank channel contains harmonics, the value of FBE obtained by using the standard calculation is affected by low spectral values between harmonics. As such, any noise present between harmonics can easily cause a mismatch between the FBEs of clean speech and noisy speech. This situation may be improved by calculating the FBEs based only on few highest values of $S_w(k)$. Moreover, for filter-bank channels whose $vd^{fb}(b)$ is below a voicing-distance threshold (i.e. estimated as voiced), only the highest values of $S_w(k)$'s whose associated $vd(k)$'s are also below the threshold are considered. The number of highest values used in the summation was set from three to six for channels 1 to 5, 6 to 10, 11 to 15, 16 to 20, respectively. Note that each FBE value was normalized by the number of components used in the summation.

3. EXPERIMENTAL RESULTS

This section presents experimental evaluation of the proposed voicing-character estimation method using real speech and of the modified calculation of filter-bank energies. First, the performance of the voicing-character estimation method is evaluated in terms of false-acceptance (FA) and false-rejection (FR) errors on a task of a binary, voiced/non-voiced, character estimation of filter-bank channels. Then, the effect of FA and FR errors is demonstrated in terms of recognition accuracy results when the voicing-character information is employed in an ASR system. Finally, the recognition performance obtained by using features derived from the modified filter-bank energies is presented.

As the true information about the voicing-character of filter-bank channels is not available, it is defined based on a-priori knowledge of clean speech signal and noise; this will be referred to as 'oracle' voicing label. Based on experimental results presented in [7], a filter-bank channel of noisy speech is assigned an oracle label *voiced* if its corresponding voicing-distance on clean speech is below the value 7 and its local-SNR is equal or above 0dB, and *non-voiced* otherwise. In experiments, a filter-bank channel is estimated as voiced when its corresponding voicing-distance value is below a voicing-threshold.

The experiments were carried out on the isolated part of the TIDigits database, which contains a total of 2486 test utterances, corrupted by various types of additive noise at global SNR equal to 20dB, 10dB, and 0dB, respectively.

3.1. Evaluation of voicing-character estimation in terms of false-acceptance and false-rejection

The results of voicing-character estimation in terms of FA and FR error rates are, for speech corrupted by white noise at various local SNRs, depicted on Figure 2. It can be seen that varying the voicing-threshold can provide the required FA or FR errors. For instance, the FA and FR errors are both approximately 5% for speech corrupted at 10dB local SNR.

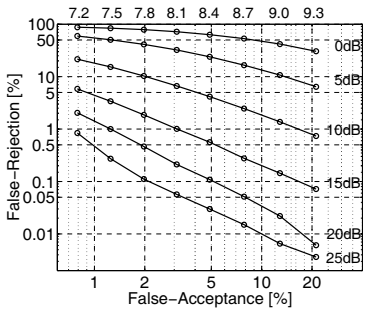


Fig. 2. *False-acceptance and false-rejection error rates of voicing-detection as a function of voicing-threshold value (above the figure) for speech corrupted by white noise at various local-SNRs.*

3.2. Evaluation when employed in a missing-feature-based ASR system

This section presents experiments when the voicing-character information for filter-bank channels is employed in a missing-feature-based ASR system. The performance of the proposed voicing-character estimation method is evaluated by comparing the recognition accuracies obtained by an ASR system that uses the information about voicing-character estimated by the proposed method and oracle voicing-information. This reflects the effect of FA and FR errors presented in the previous section on the recognition accuracy of an ASR system.

The experiments were carried out for speaker-independent digit recognition. The features used for speech recognition were obtained by filtering the logarithm filter-bank energies over the frequency dimension by the filter $H(z)=z-z^{-1}$ [8], resulting a feature vector consisting of 18 elements (denoted as FF-feature vector). In order to include dynamic spectral information, the first-order delta parameters were added. A continuous-observation left-to-right HMM with 16 states (no skip allowed) was used to model each digit. For each state, three Gaussian mixtures with diagonal covariance matrices were used. The training of HMMs was performed on clean utterances from the training set by using all the features. In recognition, marginalization-based missing-feature ASR system was employed. A static FF-feature was used only if both filter-bank channels (used for its calculation, see above) were labelled as voiced, otherwise it was marginalized. The dynamic features were used all. The voicing-labels for filter-bank channels were obtained by setting the voicing-threshold to 8.5; negligible performance differences were observed

when the threshold value was within the range from 8 to 9. For comparison, the experiments were also performed when all features were used (i.e. standard method) – this corresponds to a situation when all features would be estimated as voiced.

The experimental results are presented in Table 1. It can be seen that employment of the voicing-information obtained by the proposed method can significantly improve the recognition performance over the standard method. Indeed, the recognition performance obtained when using the voicing-labels estimated by the proposed method are very similar to using the voicing-labels obtained based on a-priori knowledge of noise. Slightly larger recognition accuracy difference in the case of Pub noise is due to speech-like content of this noise, i.e. some voiced features actually correspond to background noise. This can be improved by combining the features detected as voiced by using the feature-combination model presented in [9].

Table 1. *Recognition accuracy results obtained by an ASR system that uses voicing-labels estimated by the proposed method and by full a-priori knowledge about noise (oracle). The performance when using all features is included for comparison.*

SNR [dB]	Noise type	All Features	Voiced Features	
			Oracle	Estimated
20	White	94.0	95.8	95.9
	Factory	96.5	96.9	97.0
	F16	92.5	96.7	96.6
	Pub	95.3	96.6	95.8
10	White	77.0	87.4	87.4
	Factory	87.8	90.1	90.2
	F16	68.6	89.5	88.9
	Pub	82.4	89.6	86.1
0	White	35.7	52.7	52.9
	Factory	50.6	55.1	53.8
	F16	21.6	55.4	51.9
	Pub	50.6	63.8	56.1

3.3. Experiments with features derived from the modified FBEs calculation

This section presents experimental results when using the same estimated voicing-labels as in the previous section, however, the features derived from the modified calculation of filter-bank energies. The recognition results for speech corrupted at SNR=0dB are presented in Figure 3. It can be seen that using features based on modified calculation of FBEs provides improvement in all noisy conditions, except for the Pub noise. Note that the results for speech corrupted at higher SNRs are not presented as there was only a negligible difference (less than 0.5%) in recognition performance between using standard and modified FBEs calculation. This may be due to lower noise power – as such its presence between harmonics will have only little effect on the filter-bank energy values.

4. CONCLUSION

In this paper, we presented a simple yet effective method for the estimation of the voicing-character of speech spectra based on the comparison of the spectral shape around each peak of short-term spectra to the spectra of the frame-analysis window. This method does not require information about the fundamental frequency. The proposed method was evaluated on voiced/non-voiced character estimation of filter-bank channels. Experiments were performed for speech corrupted by various noises at various SNRs. The results of detection of voiced filter-bank channels for speech corrupted by white noise at 10dB local SNR show less than 5% FA and FR errors. The voicing-character information was incorporated in a missing-feature based ASR system. The experimental results showed very similar performance obtained by using the features estimated as voiced by the proposed method and by full a-priori knowledge of the noise. Finally, a modification to the calculation of filter-bank energies was presented and this showed further error reduction.

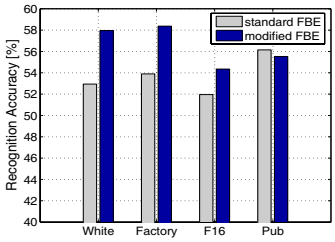


Fig. 3. *Recognition accuracy results employing the estimated voicing-labels for speech corrupted at 0dB when using features derived from the standard and modified calculation of FBEs.*

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

5. REFERENCES

[1] D. Griffin and J. Lim, "Multiband-excitation vocoder," *IEEE Trans. on Acoustic, Speech, and Signal Proc.*, vol. 36, pp. 236–243, Feb. 1988.

[2] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[3] B. Yegnanarayana, Ch. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, no. 1, pp. 1–11, Jan. 1998.

[4] P.J.B. Jackson and Ch.H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 7, pp. 713–726, Oct. 2001.

[5] 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.

[6] G. Fant, *Acoustic Theory of Speech Production*, The Hagues:Mounton, 1960.

[7] P. Jančovič and M. Kokuer, "Estimation of Voicing-Character of Speech Spectra based on Spectral Shape," *submitted to IEEE Signal Processing Letters*, 2005.

[8] 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.

[9] P. Jančovič, M. Kokuer, and F. Murtagh, "High-Likelihood Model based on Reliability Statistics for Robust Combination of Features: Application to Noisy Speech Recognition," *Eurospeech, Geneva, Switzerland*, pp. 2161–2164, 2003.