ON THE IMPORTANCE OF COMPONENTS OF THE MODULATION SPECTRUM FOR SPEAKER VERIFICATION

Sarel van Vuuren[1] *Hynek Hermansky*[1,2]

[1] Department of Electrical and Computer Engineering, Oregon Graduate Institute of Science and Technology,
PO Box 91000, Portland, OR 97291-1000, USA, [2] International Computer Science Institute, Berkeley, California, USA
sarelv@ece.ogi.edu, hynek@ece.ogi.edu

ABSTRACT

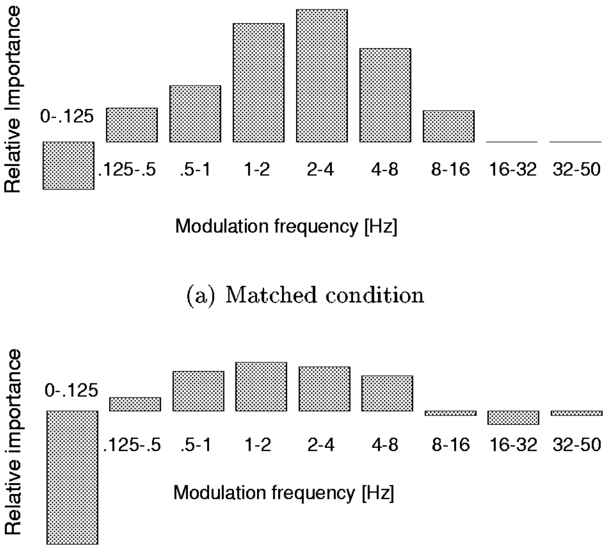
We provide an analysis of the relative importance of components of the modulation spectrum for speaker verification. The aim is to remove less relevant components and reduce system sensitivity to acoustic disturbances while improving verification accuracy. Spectral components between 0.1 Hz and 10 Hz are found to contain the most useful speaker information. We discuss this result in the context of RASTA processing and cepstral mean subtraction. When compared to cepstral mean subtraction that retains components up to 50 Hz, lowpass filtering to 10 Hz with downsampling by 75% is found to significantly improve robustness in mismatched conditions. The downsampling results in a large computational savings.

1. INTRODUCTION

Many speaker verification systems attempt to characterize a speaker using acoustic features based on *filtered* logarithmic spectral energies derived from a short-time analysis [1, 5, 2]. Spectral components of the time sequences of logarithmic spectral energies, aka. the modulation spectrum, are affected by this filtering. It is therefore of interest to determine the relative importance of the spectral components for speaker verification.

Delta processing, a polynomial regression (differentiation) spanning about 50 ms of speech [2], is often used as a filter to extract temporal information. Similarly, to suppress convolutional noise (eg. frequency characteristics of a communication channel which is additive in logarithmic spectrum or cepstrum), cepstral mean subtraction (CMS) [2], or mean subtraction (MS) as referred to in this paper, is used as a filter to suppress DC components in the time sequences of the logarithmic spectral energies. Another technique [1] that limits the frequencies present in the spectral trajectories is RelAtive SpecTrAl Processing (RASTA). Figure 2 shows the frequency responses of these filters for a 100 Hz sampling rate of the logarithmic spectral energies. The MS filter has a highpass frequency response with cut-off frequency depending on the length of the averaging window (here 0.025, 0.075 and 0.25 Hz respectively for window lengths of 30, 10 and 3 seconds). The RASTA filter has a passband of about 1 to 13 Hz. The delta polynomial (-2,-1,0,1,2) computed in a 50 ms window of speech has a passband of about 7 to 21 Hz.

Obviously filtering should be used to enhance speaker specific information while suppressing non-informative and possibly confusing information. This suggests an analysis of the relative importance of the components of the modulation spectrum for speaker verification.



(a) Matched condition

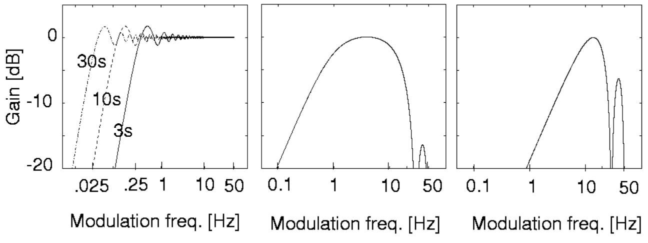
(b) Mismatched condition

Figure 1: Relative importance R of components of the modulation spectrum. Positive values indicate a decrease in verification error contributed to the inclusion of a particular modulation spectral band in the acoustic features. Results were derived on 30 second test segments (male and female) from the 1997 NIST-SRE corpus.

2. EXPERIMENTAL SETUP

2.1. Acoustic feature processing, statistical model and decision score

Acoustic features for the verification experiments are derived from a short-time analysis of the speech signal with a 32 ms analysis window advanced in 10 ms steps. Logarithmic spectral energies are computed from the squared magnitude FFT using a triangular integration window in a manner similar to that of the computation of Mel-Frequency Cepstral Coefficients [2]. The 19 spectral energies falling within the range of 200 to 3400 Hz are retained. Each spectral energy is further processed by one or more FIR filters. The effect of various choices for these filters



(a) MS (b) RASTA (c) DELTA

Figure 2: Frequency responses of various filters in the modulation spectral domain.

are analyzed in Section 3. After filtering, vectors derived from non speech frames are dropped using an adaptive energy-based detector that discards frames with energies below the estimated noise floor in the signal. The vectors are decorrelated using a Karhunen Loéve transformation computed on the training data. Parameterizations were optimized on data from the 1996 NIST Speaker Recognition Evaluation (NIST-SRE) [8].

Speaker dependent (SD) models are trained on speaker data by MAP reestimation of the parameters in a speaker independent (SI) model [6]. The SI model (a mixture of Gaussians with either 128 or 256 components) is trained using the EM-algorithm on a population of 40 male and 40 female speakers different from those used in training and testing. Parameters are initialized using the LBG algorithm with iterative cluster splitting. This system [7] performed competitively in the 1998 NIST-SRE.

To evaluate the claim of a specific speaker having produced given test vectors, the likelihood of that speaker's SD model normalized by the likelihood of the SI model is compared to a speaker independent threshold[1]. Conveniently, likelihoods for the SD and SI models are accumulated using only the five best scoring components identified for the SI model on the test vectors [6]. Verification performance is evaluated using the Equal Error Rate (EER) and a decision cost function (DCF) used in the NIST-SRE.

2.2. Speech data

Results are based on continuous telephone speech sampled at 8 kHz from the Switchboard-2 phase 1 and 2 corpora as used in the 1997 and 1998 NIST Speaker Recognition Evaluations [8]. The SI model is trained on 1997 NIST-SRE data, using 40 male and 40 female speakers. Each SD model is trained on data from two 1 minute segments of speech, each from a different recording session (2-session).

Testing is performed separately for nominal durations of test speech of 3, 10 and 30 seconds in both matched and mismatched communication channels. In the matched condition, handset type (electret or carbon button) and telephone number are the same for both training and test utterances; for the mismatched condition, handset type and telephone number are different[2]. Table 1 summarizes the test conditions used in this paper as pertaining to these corpora. Note that tests conducted on the 1997 NIST-SRE corpus use only a feature vector A derived as the output of filter H_1 in Figure 7 while tests conducted on the 1998 NIST-SRE corpus use a feature vector (A,B) derived from both filter outputs H_1 and H_2.

CORPUS	1997 NIST-SRE	1998 NIST-SRE
Features (Figure 7)	A (static)	A, B (dynamic)
# GMM components	128	256
# of target speakers	199 (67 male)	500 (250 male)
Test segment duration (seconds)	3, 10, 30	30
# of test segments per segment duration	5889 (2623 male)	5000 (2500 male)
# of target tests per test segment	10	10
Total # of tests per segment duration	58890	50000

Table 1: Statistics of Switchboard-2 phase 1 and 2 corpora as used for testing in this paper.

3. RESULTS

3.1. Relative importance of components of the modulation spectrum

Figure 1 depicts the relative importance of different components of the modulation spectrum for SV on the 1997 NIST-SRE corpus. A positive value for a band reflects a relative reduction in verification error due to the inclusion into the acoustic features of the components of the modulation spectrum within that band. Components around 2 to 4 Hz are seen to be relatively more important to reducing the error rate. This is interesting since dominant rates of change in the logarithmic power spectrum of speech have been estimated at around 2 to 4 times per second [3, 4].

The charts were derived from verification error rates obtained by bandpass filtering in the modulation spectral domain with different low f_l and high f_h frequency cut-offs ranging from 0 to 50 Hz on a logarithmically spaced grid[3]. In each case the Equal Error Rate $e(f_l, f_h)$ as a function of low and high cut-offs is computed using features with modulation spectral components from that particular passband only $f_l \leq f < f_h$. As an example, Figure 3 depicts the grid generated by cut-offs at 2, 4, 8, 16, and 32 Hz. To derive the charts, normalized differentials with respect to each of the low and high cut-off frequencies for the surface described by the error function $e(f_l, f_h)$ are averaged[4]. The

[1] Speaker or handset specific normalizations are not used.

[2] These distinctions were assessed using telephone number and handset type labels distributed by NIST.

[3] Filters were designed to have frequency responses with similar shape on the grid with sharp cross-overs and 50dB attenuation in the stop band.

[4] This procedure is similar to one described in [4] except for the normalization applied here.

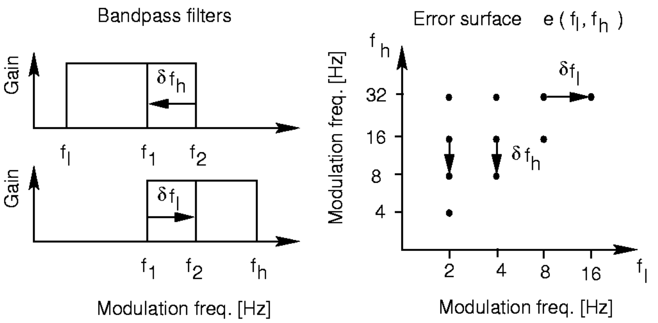


Figure 3: Grid for evaluating the importance of components of the modulation spectrum for SV.

average relative importance $R(f_1, f_2)$ of the spectral components between f_1 and f_2 is estimated as the average of the normalized partial differentials evaluated in the region $f_1 \leq f < f_2$. That is

$$R(f_1,f_2) = \frac{1}{n}\left[\sum_{f_i<f_2}\frac{e(f_i,f_2)-e(f_i,f_1)}{e(f_i,f_1)}\right.$$

$$\left. + \sum_{f_j>f_1}\frac{e(f_1,f_j)-e(f_2,f_j)}{e(f_2,f_j)}\right], \qquad (1)$$

where n is equal to the number of terms in the summation. Figure 3 depicts this computation for $R(8, 16)$.

Note that the computation amounts to estimating an averaged gradient for the logarithmic error surface since for a band $\mathcal{F} = [f_1, f_2]$

$$\left.\frac{\partial \log e(f)}{\partial f}\right|_{f\in\mathcal{F}} = \left.\frac{1}{e(f)}\frac{\partial e(f)}{\partial f}\right|_{f\in\mathcal{F}} \approx \left.\frac{e(f_2)-e(f_1)}{e(f_0)(f_2-f_1)}\right|_{f_0\in\mathcal{F}}$$

as evaluated for the low and high cut-off frequencies separately.

A positive value for the average relative importance $R(f_1, f_2)$ reflects a relative reduction in verification error due to the inclusion into the acoustic features of the components of the modulation spectrum within the band $f_1 \leq f < f_2$. It should be noted that the measure $R(f_1, f_2)$ used here only provides an indication of importance of a band and as such does not provide information on the inter-dependence (eg. correlation) of different spectral components for SV.

3.2. Effect of highpass filtering

As can be seen in Figure 1 inclusion of components of the modulation spectrum below about 0.125 Hz increases error rate, while inclusion of components above about 0.125 Hz leads to a decrease in error rate. This suggests that the highpass cut-off of 1 Hz (see Figure 2) used with RASTA filtering in ASR should be lowered for SV. Figures 4 and 5 confirms this observation. The Figures show EER as a function of cut-off frequency for highpass filtering ($[f_l, 50]$ Hz). The Figures show that reducing the highpass cut-off generally reduces the error rate, with an optimum reached at a cut-off frequency close to that of MS.

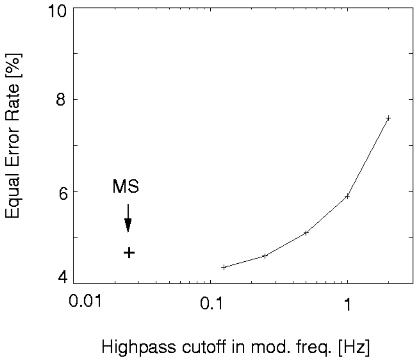


Figure 4: Matched condition. EER versus highpass cut-off for verification of 30 second test segments from the 1997 NIST-SRE corpus.

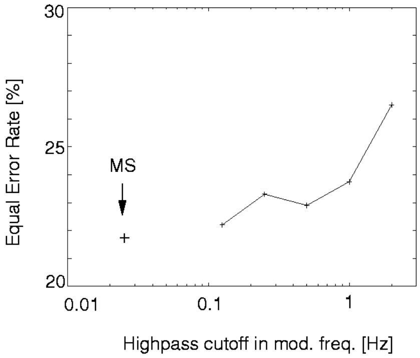


Figure 5: Mismatched condition. EER versus highpass cut-off for verification of 30 second test segments from the 1997 NIST-SRE corpus.

3.3. Effect of lowpass filtering

In the matched condition, inclusion of components above 16 Hz increases the error rate, while in the mismatched condition inclusion of components as low as 8 Hz increases the error rate. This indicates that higher modulation frequencies may not be important for SV and that removing them may improve performance. Figure 6 justifies this claim with a comparison of a lowpass system with $[f_l, f_h] = [0.0125, f_h]$. The Figure shows EER for matched and mismatched conditions and for testing with 30 second segments from the 1997 NIST-SRE. In the mismatched condition a lowpass system at 10 Hz results in a relative reduction in EER of more than 14% while in the matched condition the lowpass system results in a relative reduction in EER of more than 8%. It appears that lowpass filtering to about 10 Hz may help to alleviate communication channel mismatch above and beyond the removal of convolutative mismatch by MS.

3.4. Effect of lowpass filtering and downsampling

In the previous sections it was observed that components of the modulation spectrum at relatively low frequencies (down to 0.125 Hz) contain useful speaker information and should not be removed. It was also observed that the removal of higher frequencies (above 10 Hz) reduces verifi-

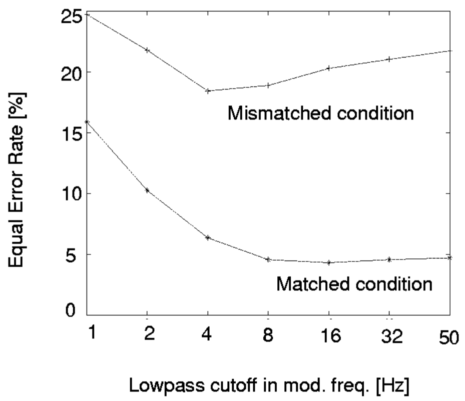


Figure 6: EER versus lowpass cut-off for verification of 30 second test segments from the 1997 NIST-SRE corpus.

cation error in the mismatched condition. Based on these observations it appears reasonable to process the logarithmic spectral energies with a bandpass filter that preserves modulation frequencies between about 0.125 and 10 Hz. We investigate the usefulness of such filtering here by using a combination of MS and lowpass filtering to 10 Hz. We downsample the time sequences since the lowpass filtering removes their higher modulation frequency components. Figure 7 depicts the proposed processing for deriving features. Static features are derived by applying MS and a 101-tap lowpass FIR filter with cut-off at 10 Hz to the time sequences of logarithmic spectral energies (filter H_1). Dynamic features are derived by applying a delta filter (50 ms window) to the static features (filter H_2). The Figure shows the composite frequency responses at outputs A and B of these filters. They are to be compared to the frequency responses in Figure 2. Both the static and dynamic features are downsampled from 100 Hz to 25 Hz, allowing a computational savings when modeling and scoring the speech.

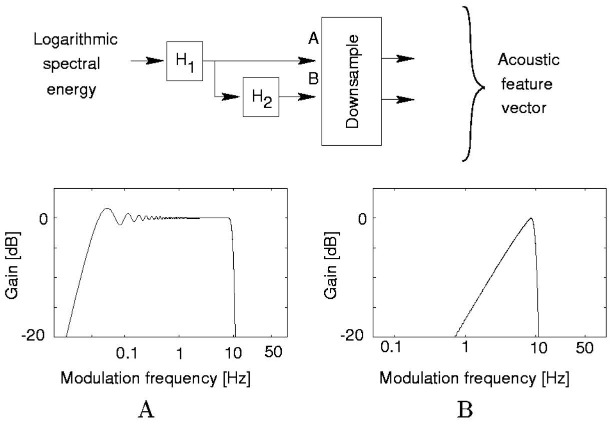


Figure 7: System and filter frequency responses for deriving acoustic feature vectors from logarithmic spectral energies.

Verification results are compared for two systems on the 30 second test segments (male and female) from the 1998 NIST-SRE corpus. The baseline system uses MS but does not use lowpass filtering or downsampling ([0.025, 50] Hz, sampled at 100 Hz). The test system uses MS, lowpass filtering and downsampling ([0.025, 10] Hz, sampled at 25 Hz). Compared to the baseline system, the test system results in a 10.7% relative reduction in the DCF and a 13.2% relative reduction in the EER in the mismatched condition. In the matched condition the test system results in a 1.4% relative reduction in the DCF and a 3.2% relative increase in EER. Futhermore, the downsampling used in the test system results in a 75% computational savings since only the downsampled features are modeled and scored.

4. CONCLUSION

We conclude that spectral components between 0.1 Hz and 10 Hz contain the most useful speaker information. We conclude that a RASTA-type processing may be useful for speaker verification, provided that frequency components below 1 Hz are retained. We show that lowpass filtering to about 10 Hz and downsampling to 25 Hz preserve salient speaker information while improving robustness. On the Switchboard corpus, this processing results in about a 10% relative reduction in error when there is a mismatch of the communication channel between training and testing.

Acknowledgments

This work was supported by DoD grant MDA-904-97-1-0098.

5. REFERENCES

1. H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.
2. S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. ASSP*, vol. 29, pp. 254–272, April 1981.
3. S. van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters," in *Proc. EUROSPEECH*, (Rodos, Greece), pp. 409–412, 1997.
4. N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," in *Proc. EUROSPEECH*, (Rhodes, Greece), pp. 1097–1101, 1997.
5. C. Nadeu, P. Pachés-Leal, and B. H. Juang, "Filtering of time sequences of spectral parameters for speech recognition," *Speech Communication*, vol. 22, pp. 315–332, 1997.
6. D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. EUROSPEECH*, (Rhodes, Greece), pp. 963–970, 1997.
7. S. van Vuuren and H. Hermansky, "!MESS: A modular, efficient speaker verification system," in *RLA2C*, (Avignon, France), pp. 189–201, April 1998.
8. *Speaker Recognition Workshop Notebook*, NIST, 1998. NIST speaker recognition evaluation on the Switchboard Corpus.