On the Use of Missing Feature Theory with Cepstral Features

Juha Häkkinen[1] and Hemmo Haverinen[2]
[1] Nokia Mobile Phones, Tampere, Finland
[2] Nokia Research Center, Tampere, Finland
juha.m.hakkinen@nokia.com and hemmo.haverinen@nokia.com

Abstract

Missing feature theory has been proposed as a solution for either ignoring or compensating the unreliable components of feature vectors corrupted mainly by bandlimited background noise. Since the corruption often occurs in the frequency domain, and it is smeared by the discrete cosine transform used to obtain cepstral features, algorithms utilizing the missing feature theory are usually restricted to spectral features. In many cases cepstral features might be preferable. We propose an algorithm for performing the missing feature operations in the frequency domain while utilizing standard Mel-cepstral features. The algorithm is based on transforming the cepstral difference operation into the Mel-spectrum domain, weighting (marginalizing) it there, and transforming it back into the cepstrum domain. The algorithm was tested in a German connected digit recognition system. Experimental results are shown first with artificially corrupted speech and then with speech corrupted using more realistic noise. While the results with data having clearly localized frequency domain corruption prove the viability of the proposed method, the results with car noise with more wideband-like characteristics were disappointing.

1. Introduction

Voice dialing is nowadays probably one of the most widely spread speech recognition applications. It is typically used in the car environment. Car noise is characterized by high intensity and relatively wide bandwidth, but it is fairly stationary. Several techniques have been proposed for coping with stationary wideband background noise. However, in some cases, the signal-to-noise ratio (SNR) of certain frequency bands of the speech signal may be so low that virtually no information is left. We observed this kind of a situation when a fairly dramatic reduction of the performance of an otherwise robust connected digit recognizer was caused by a severely corrupted frequency region of the input speech signal.

One approach for handling bandlimited noise is so-called sub-band speech recognition (see, e.g., [1] or [2]). The basic principle is to divide the speech signal into sub-bands and employ independent speech recognizers in each sub-band. Decision logic is needed for the combination of the recognition results.

Missing Feature Theory (MFT) has also been proposed as an alternative solution for handling situations where the incoming data is partially corrupted. An overview of MFT is given in [3]. In the most straightforward approach, which is called marginalization, the summation (or integration) is performed over the difference of the reliable feature and model vector components, while ignoring the unreliable ones. Measuring the reliability of feature vector components is a challenging topic, but using an on-line estimate of the SNR has been the most widely used method. Marginalization typically involves hard-decisions; a component is labeled either reliable or unreliable. Soft-decision techniques have been applied to MFT, e.g., in [4], where some performance gain was obtained. Data imputation is a more elaborate approach to MFT, where the missing data is inferred from the neighboring feature vector components by means of interpolation techniques.

The MFT techniques in general suffer from the need to perform the operations in the feature vector domain, which is not necessarily where the corruption of the data takes place. For example, bandlimited noise in the frequency domain is smeared by the discrete cosine transform (DCT) used to obtain widely used cepstral feature vectors. This has precluded the use of cepstral feature vectors with the MFT techniques. Marginalization in the cepstrum domain does not really make much sense since noise seldom affects only certain cepstral coefficients.

The main objective of the proposed approach is to experiment with the advances made in the domain of missing feature theory when using Mel-frequency cepstral coefficient (MFCC) feature vectors, and feature vector normalization, in a connected digit recognition system. Our speech recognition framework has been presented in [5]. We assume spectrally localized corruption of speech signals. We propose a straightforward approach for canceling the effect of corrupted frequency regions while operating on MFCC features. Although we concentrate on marginalization, implementation of data imputation techniques is also possible. The proposed approach is called Cepstral Distance Weighting (CDW) algorithm.

The paper is organized as follows. We first present our speech recognition framework and the basic CDW algorithm, and how it relates to missing feature theory. Second, we discuss the practical implementation of the algorithm. Then we present our experimental results and finally conclude the paper with some discussion on potential directions for further research on the subject.

2. Cepstral Distance Weighting Algorithm

2.1 Speech Recognition Framework

We use the standard Hidden Markov Model (HMM) speech recognition framework and a Mel-Frequency Cepstral Coefficient (MFCC) based front-end with 39 components (13 static, 13 delta, and 13 delta-delta components). The front-end uses an 8 kHz sampling frequency and a 10 ms frame period. Gaussian mixture density HMMs are used by the back-end. The Viterbi algorithm is implemented using the token passing scheme. The basic system is described in more detail in [5].

For the sake of the simplicity of presentation of the algorithm, we are now assuming the use of only one single-state, single-mixture HMM. Extension into multiple models, multiple states per model and multiple mixtures per state is

straightforward. The log-likelihood (Mahalanobis distance) of the input feature vector **f** given the model $\lambda = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ (where $\boldsymbol{\mu}$ denotes the mean vector and $\boldsymbol{\Sigma}$ the covariance matrix) is calculated as

$$D(\lambda, \mathbf{f}) = c(\boldsymbol{\Sigma}) - \tfrac{1}{2}\mathbf{d}^T \boldsymbol{\Sigma}^{-1}\mathbf{d} \quad (\mathbf{d} = \mathbf{f} - \boldsymbol{\mu}), \qquad (1)$$

where $c(\boldsymbol{\Sigma})$ represents the constant term of the log-likelihood estimation formula.

In order to improve the robustness of the recognizer, recursive feature vector normalization is used as in [5]. The normalization algorithm is described in more detail in [6]. To summarize, the main objective is to increase the robustness of the feature vectors by normalizing the short-term mean and standard deviation of each component to zero and unity, respectively.

2.2 CDW Algorithm Using MFCC Features

In the standard MFT approach, where spectral features are used, the effect of a single corrupted frequency component is canceled by marginalization, i.e., the Mahalanobis distance is evaluated only on the reliable frequency components. Marginalization can be implemented in our framework by the addition of a weighting matrix **W**, which is a diagonal matrix with one weight w_i (typically between zero and one, or exactly zero or one as in the case of marginalization) for each frequency component *i*. Delta and delta-delta components have their own frequency-dependent weights in a similar fashion. The Mahalanobis distance is now evaluated by

$$D^{MFT}(\lambda, \mathbf{f}) = c(\boldsymbol{\Sigma}) - \tfrac{1}{2}\mathbf{d}^T \mathbf{W}^T \boldsymbol{\Sigma}^{-1}\mathbf{W}\mathbf{d}\,. \qquad (2)$$

The main problem of the classical MFT is that if cepstral features were used, spectrally localized corruption of the speech signal would be smeared by the discrete cosine transformation matrix[1] **C**, and the benefit of MFT would be lost.

A simple modification of (2) enables us to use spectrum domain marginalization with cepstral features, as shown by

$$D^{CDW}(\lambda, \mathbf{f}) = c(\boldsymbol{\Sigma}) - \tfrac{1}{2}\mathbf{d}^T \mathbf{C}\mathbf{W}^T \mathbf{C}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{C}\mathbf{W}\mathbf{C}^{-1}\mathbf{d}\,, \qquad (3)$$

where $\mathbf{C}^{-1} = \mathbf{C}^T$. The best interpretation of (3) is obtained by observing the right-hand side of the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$. The cepstral difference vector **d** is first reverted back to the log-Mel-spectrum domain (a smoothed version is obtained if truncation was applied after the forward DCT), where marginalization, or some other kind of weighting, is performed. After this, the weighted distance vector is transformed back to cepstrum domain. A few additional observations that can be made from (3):

1. Inverse DCT of $\boldsymbol{\mu}$ reverts the model mean vector into spectrum domain which enables the use of marginalization.
2. Since the weighting operates on the difference vector **d**, there is no need to obtain an accurate "clean" estimate of either the feature vector or the model mean in the noisy environment. These are typically required by feature enhancement and model compensation schemes, respectively.
3. Since delta and delta-delta operations are linear, it is possible to obtain spectrum domain estimates for each of them by means of the inverse DCT, which enables the use of CDW.

We have also experimented with using Minimum Classification Error training of the relationship between the SNR and the weights. Our early results did not produce any significant improvements, which is probably due to the apparent hard-decision nature of the weighting process. This is illustrated in Figure 4 in the Experiments section of the paper.

2.3 CDW Algorithm Using Feature Vector Normalization

The use of feature vector normalization changes the interpretation of (3). By applying the inverse DCT to **d** after the normalization operations (normalization is a lossy process), we are no longer able to revert back to the Mel-spectrum domain. However, we make the assumption that normalization does not significantly change the localization of noise in the frequency domain. Experimental results presented later in this paper verify this assumption at least to some degree.

3. Implementation of the CDW Algorithm

The implementation of (3) in a practical system with hundreds or thousands of mixture vectors is next to impossible because of the extensive matrix operations. If the weights are changed less frequently than the normal frame rate (i.e., decimation is used), some performance savings can be obtained.

When MFCC features are used, we can save a Mel-domain version of the input feature vector that removes the need to perform the inverse DCT on it. When feature vector normalization is used, it is imperative to first calculate the normalized feature vector and then obtain the Mel-domain version by means of the inverse DCT.

Instead of performing (3) on distance vectors corresponding to every mixture, the feature vector and the mixture mean vectors should be weighted separately and saved for later use. The complexity of the weighting operation on the mixture means (by far the most expensive part) is then reduced roughly according to the decimation rate. Decimation may be a good idea also from the algorithm performance viewpoint as too rapidly changing weights might cause problems.

3.1 Weight Estimation Algorithm

The idea is to weight the Mahalanobis distance according to SNR of the Mel-bands. First, we assume that noise is stationary. The update of the weight parameters is only done once in the beginning of the utterance, after a reasonable noise estimate has been obtained. The noise estimate for each of the sub-bands is based on the average of the 20 first frames. The speech power is estimated from the previous utterance, i.e., we assume that the power of the speech between two successive utterances is changing rather little. The SNR is then calculated based on these speech and noise estimates.

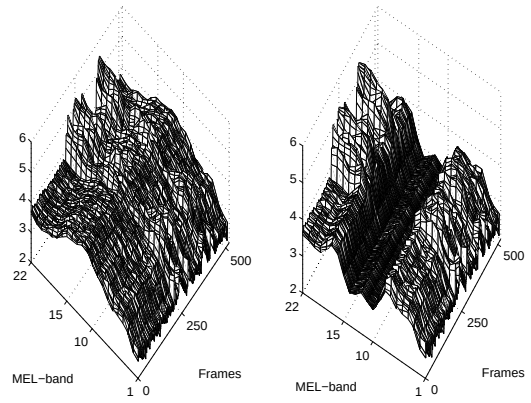
The separation between unreliable and reliable sub-bands is done based on the difference between the highest and lowest SNR of the Mel-bands. If the difference is small, CDW is disabled. If the difference is above the threshold, CDW is enabled and the weights are linearly interpolated between the minimum and maximum values of the weight, i.e., between 0.3 and 1.0. The weight zero (corresponding to marginalization) was actually found to be too extreme.

[1] **C** is actually block-diagonal with the same transformation sub-matrix for static, delta, and delta-delta components.

One test having more dynamic changing of the weights was done. The weights were updated every 5th frame according to SNR. Since this produced only small improvements in the very low SNR ranges (around -10 dB) and the system became more complex, this was not included in the tests.

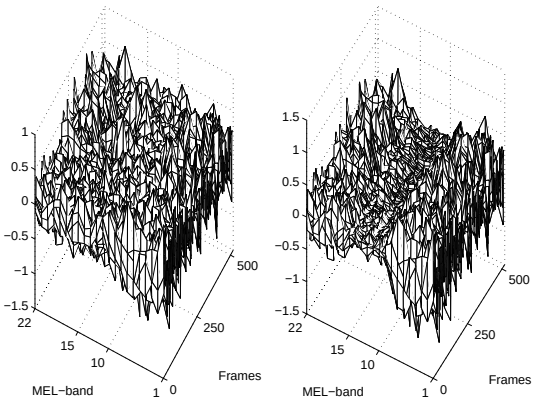
4. Experiments

Recognition tests were performed on German digit sequences corrupted both with artificially generated and recorded car noise. The Mel-spectra of the corrupted utterances are shown in Figure 1. The spectra have been calculated by taking the inverse DCT of the unnormalized cepstral features.



a) *b)*

Figure 1: The Mel-spectrum of the noisy speech sample. *a)* The speech is corrupted by car noise (mostly around sub-band 10). *b)* Sub-bands 10…14 are set to a constant value in the front-end.



a) *b)*

Figure 2: The Mel-spectrum of the noisy speech sample (Inverse DCT of normalized cepstral features). *a)* The speech is corrupted by car noise (mostly around sub-band 10). *b)* Sub-bands 10…14 are set to constant in the front-end.

The noise robustness of the recognition system is improved by using feature vector normalization and hence we wanted to utilize normalization also with the weighting algorithm. Since we do not have reasonable inverse normalization procedure, the weighting has to be done for spectral features that correspond to the normalized cepstral features.

Figure 2 illustrates how feature vector normalization, which is done in the cepstrum domain, changes the speech spectrum. The speech sample used in these graphs is the same as in Figure 1. The spectrum is obtained by taking the inverse DCT of the normalized cepstral features. The corruption is visible in the right-hand side figure, between sub-bands 10 and 14, also after feature vector normalization, although the dynamic range of the spectrum is significantly reduced.

4.1 Results with Artificially Corrupted Mel-bands

In order to show that the CDW algorithm works in practice, a simple test was carried out, where the number of artificially corrupted sub-bands was gradually increased. The actual speech data was recorded in a quiet environment and the corruption was done in the front-end simply by setting some of the Mel-bands to a small constant value.

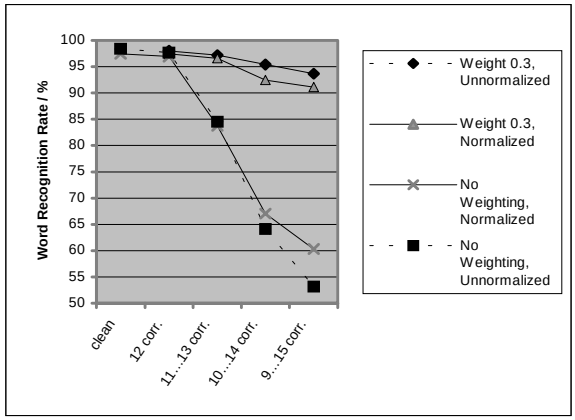


Figure 3: Word recognition rate with and without the weighting algorithm when the number of artificially corrupted sub-bands was increased (the indices of the corrupted bands are shown). Normalized and unnormalized features are also compared.

CDW was done only for those sub-bands that were corrupted. First, only the 12th sub-band (around 1.3 kHz) was corrupted and the recognition rates were computed both with and without weighting. The tests were continued by corrupting more and more sub-bands and checking the rates after that with and without weighting. Significant improvements were obtained by using the CDW algorithm. When the number of corrupted sub-bands increases, the recognition rate degrades much faster if no weighting is used than if a small weight is given to the corrupted sub-bands during the recognition phase.

When testing with normalized features, the recognition rate improvements were quite similar to the ones obtained using unnormalized features. The conclusion is that the normalization process does not disturb the CDW algorithm significantly. Since it makes the system more robust to many other kinds of noises, feature vector normalization is used for the car noise experiments presented later in this paper.

The recognition performance is presented in Figure 4 as a function of the weight parameter value. The sub-bands 9…15 are corrupted in the front-end and the weight of these sub-bands is changed between 0.0 and 1.0. Two different levels of corruption were tested. First, the Mel-bands were set to a small constant (=3.0) as before, and then to zero. It can be seen that

when a small weight is used both unnormalized and normalized features work well with both corruption types. The performance of the normalized features (dashed lines) does not depend at all on the level of corruption in this test. When the weights are larger than about 0.1, and the corruption is severe, the performance of the unnormalized features degrades rapidly.

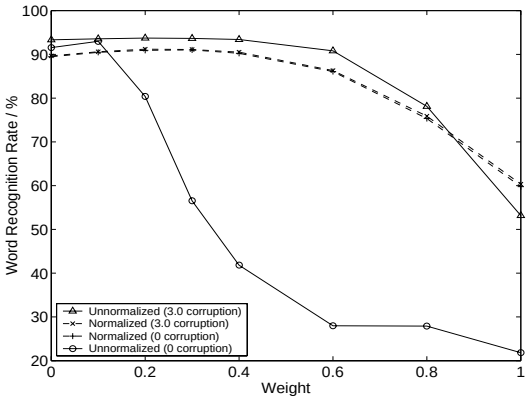


Figure 4: Word recognition rates using different weight values when sub-bands from 9 to 15 were corrupted.

4.2 Results with Car Noise

The performance of the CDW algorithm was also evaluated with car noise simulations. Feature vector normalization was used in these tests. The database contains digit strings that were recorded in a car simulator where generated car noise was played in the background using the multiple loudspeakers generating a realistic sound field. This database is suitable for testing MFT-based techniques since the corruption of the frequency region around 1 kHz is emphasized (see Figure 1a). However, the corrupted part of the spectrum does not change much since the noise is quasi-stationary.

Table 1: The results from the first car noise tests.

	No Weighting	Weighting
Word rate	85.44	89.82
String rate	64.36	73.96
Insertions	150	94
Deletions	104	79
Substitutions	816	575

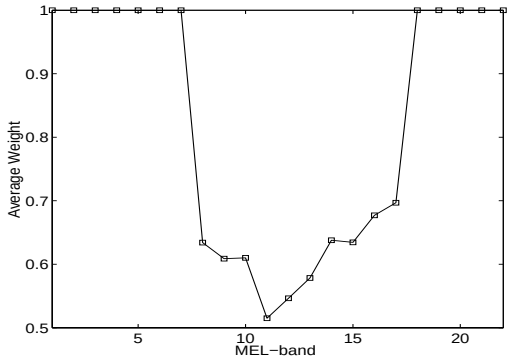


Figure 5: Average weights used for the utterances recorded in the car simulator.

The results of the car noise tests are summarized in Table 1 which shows the substantial improvement due to the CDW algorithm. Figure 5 depicts the average weights used in the tests.

The system was also tested with another kind of car noise that had a more uniform spectrum without big differences between the average SNRs of the Mel-bands. The system was not able to improve recognition rates in this kind of noise.

5. Conclusions and Discussion

Missing feature theory based techniques have been shown to improve the recognition performance when partial corruption of the speech spectrum occurs. Since missing features are usually taken into account in the spectrum domain, the recognition is typically done using spectral features. However, the cepstrum is usually regarded as a more robust domain for feature extraction.

We proposed a novel method for suppressing unreliable frequencies in the log-likelihood evaluation phase, in a speech recognizer that uses cepstral features. The method is shown to significantly improve the recognition rates when narrowband corruption of speech occurs. We failed to gain any improvements by using on-line updating of the weights (similarly to spectrographic missing data masks). This may be due to lack of tuning of the thresholds. Although discriminative training of the weights did not yield any improvements, it might improve the performance of the frequency masks by means of finding optimum thresholds. The SNR estimation could also be performed after normalization and inverse transformation back into the Mel-domain, which would provide more accurate information.

There is most likely more potential in the proposed approach, which would merit further research. Probably the best practical set-up would be to include it in a robust speech recognizer as a supplementary means for coping with narrowband noise and distortion.

6. References

[1] H. Bourlard, S. Dupont, “A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands”, In Proc. ICSLP, pp. 426-429, 1996.

[2] H. Hermansky, S. Tibrewala, M. Pavel, “Towards ASR on Partially Corrupted Speech”, In Proc. ICSLP, pp.462-465, 1996.

[3] M. Cooke, P. Green, L. Josifovski, A. Vizinho, "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data", Speech Communication, Volume 34, Issue 3, June 2001.

[4] J. Barker, L. Josifovski, M. Cooke, P. Green, “Soft Decisions in Missing Data Techniques for Robust Automatic Speech Recognition”, In Proc. ICSLP, Sydney, Australia, pp. 373-376, 2000.

[5] J. Tang, J. Häkkinen, I. Kiss, "Improved Post-processing for Noise Robust Connected Digit Recognition", in Proc. International Workshop on Hands-Free Speech Communication, Kyoto, Japan, April, 2001.

[6] O. Viikki, D. Bye, K. Laurila, "A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise", In Proc. ICASSP, Seattle, WA, USA, pp. 733-736, 1998.