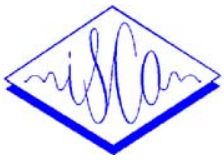
6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# OPTIMIZATION ALGORITHMS FOR ESTIMATING MODULATION SPECTRUM DOMAIN FILTERS

*Pau Pachès-Leal*†,‡, *Richard C. Rose* †, *and Climent Nadeu* ‡
†*AT&T Labs-Research, Florham Park, NJ, USA,* ‡*Univ. Politècnica de Catalunya, Barcelona, Spain*



## ABSTRACT

The goal of the work described in this paper is to develop and evaluate procedures for automatic estimation of modulation spectrum filters to compensate for distortions in the modulation spectrum domain. The modulation spectrum (MS) is often used to describe the time sequence of spectral parameters (TSSPs) that are derived from the speech waveform, and is thought to be a good representation of many sources of variability in speech. These procedures will be used in the context of automatic speech recognition (ASR) applications where there is likely to be a significant mismatch in the MS characteristics that exist for system training and evaluation. Results are presented describing application of the algorithm to one task involving an artificially introduced MS distortion and to another task involving differences in speaking styles for training and testing. It is shown in the paper that these techniques are able to compensate for the effects of artificially introduced distortions that appear in testing. It is also shown that a small degree of compensation is obtained for speaking style mismatch, and this result is compared with the measured effects of the speaking style differences in the MS domain.

## 1. INTRODUCTION

The time sequence of spectral vectors derived from the speech signal can be represented by the Modulation Spectrum (MS) [5]. It has been proposed for use in many applications. These include characterizing multipath distortions occurring in reverberant environments [2], describing the effects of channel distortion and spectral estimation errors in ASR [5], and describing the effects of varying speaking styles for ASR [5]. When used as a representation of the long–term averaged spectrum or cepstrum parameters in ASR, the MS is defined as the power spectrum of the time sequence of the feature vectors that are input to the recognizer. In speech recognition feature extraction, MS filters have been designed for the purpose of selectively removing those portions of the modulation spectrum representing noise or channel distortions, and retaining the portion of the MS containing speech [5].

MS filters are commonly applied to filtering the sequence of cepstrum coefficients and also to computing the cepstrum difference dynamic coefficients. It is important to note that the MS filters used for both the cepstrum and dynamic cepstrum coefficients are generally obtained empirically. As a result, it is often the case that MS filters designed to optimize performance for one task under a given set of conditions prove to be suboptimal when applied to another task. The automatic procedure presented here for estimating modulation spectrum filters is an attempt to improve speech recognition performance under highly mismatched recognition / training scenarios.

*This research was conducted at AT&T Shannon Laboratory as part of P. Pachès-Leal's Ph.D. thesis with the UPC

The paper is organized as follows. Section 2. describes the modulation compensation algorithm (MCA) and discusses implementation issues relating to the algorithm. In Section 3., the algorithm is implemented on an artificially introduced distortion applied to the test data. Finally, in Section 4., the issue of automatic compensation for speaking rate mis–match using the MCA algorithm is investigated.

## 2. MODULATION COMPENSATION ALGORITHM

The modulation compensation algorithm (MCA) estimates a set of filter coefficients to maximize the likelihood of the filtered observation sequence with respect to a given HMM model. The scenario under which the algorithm is applied is illustrated by the block diagram in Figure 1. The underlying assumption in this scenario is that conditions that might affect the MS characteristics of speech during recognition may not be present during HMM model training. Discussion of the MCA algorithm is presented in this section in two parts. First, it is introduced as an extension of a class of techniques developed by Chengalvarayan and Deng for simultaneous estimation of HMM and observation sequence filter parameters [1]. Second, the algorithm is described in detail, along with algorithmic issues relating to the optimization criterion and the form of the filters that are used in the algorithm.

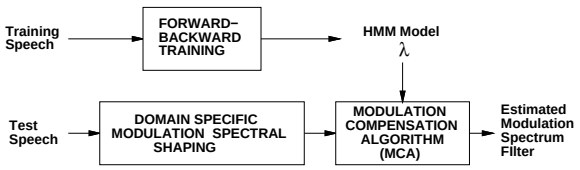



**Figure 1. Modulation compensation algorithm (MCA) applied where modulation spectrum distortions may be introduced in test conditions.**

In [1], a class of techniques for simultaneous estimation of HMM and observation sequence filter parameters was developed. This was used only for obtaining difference cepstrum parameters from the absolute cepstrum. The components of a $D$ dimensional dynamic cepstrum vector $\mathcal{Y}_t = y_1, \ldots, y_D$ were computed from a static vector $\mathcal{X}_t$ at time $t$ according to

$$\mathcal{Y}_t = \sum_{k=-b}^{f} \omega_{k,i,m} \mathcal{X}_{t+k}, \qquad 1 \le t \le T \tag{1}$$

where the "modulation" filter coefficients $\omega_{k,i,m}$ are dependent on state $i$ and Gaussian mixture component $m$. Equation 1 was introduced into the model reestimation equations for continuous Gaussian observation density HMMs and simultaneous reestimation of the filter coefficients and the HMM model parameters was performed.

10.21437/Eurospeech.1999-26

Wellekens proposed a more constrained modulation spectrum filter reestimation procedure aimed at optimizing only the cut–off frequency of the MS filters [7].

Simultaneous estimation of these parameters has the desirable effect of providing a closer coupling between model estimation and feature analysis in speech recognition. However, because of the coupling to model estimation, it is unlikely that the spectral characteristics of the filter parameters estimated in this way will actually reflect any meaningful structure associated with the MS of speech. The goal in this work is to estimate the MS filter parameters separate from the HMM model facilitating the scenario depicted by the block diagram in Figure 1. Since the model remains fixed, MS mismatch between utterances used to train the model and the utterances used for testing can be reduced. The MCA algorithm and its mathematical properties are described in more detail below.

The algorithm estimates a filter that, when applied to the absolute cepstrum coefficients, increases the likelihood of the filtered data, $\mathcal{Y}^{\omega}$, with respect to the HMM model, $\lambda$. This in theory could be accomplished by enumerating an ensemble of MS filters $\Omega$ and choosing the most likely filter in the ensemble,

$$\hat{\vec{\omega}} = \arg\max_{\vec{\omega}\in\Omega} P(\mathcal{Y}^{\omega}|\vec{\omega},\lambda) \ . \tag{2}$$

However, in practice it is very difficult to specify a manageable ensemble of filters that would be suitable for representing an arbitrary set of modulation spectrum distortions and it is not clear that a maximum likelihood criterion would be suitable for selecting the optimum MS filter from this ensemble. The algorithm described here assumes a finite impulse response filter, whose length must be chosen, and estimates the parameters of this filter using the expectation maximization (EM) algorithm. It can be applied to filtering the absolute cepstrum features or to obtaining the filter parameters used to compute the dynamic features from the absolute cepstrum. The following discussion will refer specifically to the former case.

Given an initial HMM model, $\lambda$, and an initial *length* for the MS filter, $\vec{\omega}$, this EM based algorithm maximizes the expected value of the log of the filtered data likelihood, $P(\mathcal{Y}^{\omega}|\vec{\omega},\lambda)$, with respect to $\vec{\omega}$. The data to which the filter $\vec{\omega}$ is applied can be the unfiltered absolute data, $\mathcal{X}_t$, or the absolute data filtered with an initial filter. In the latter case, the overall filter that needs to be applied to the test data so as to reduce the MS mis–match between the unfiltered training data and the test data is the convolution of the initial filter and the filter optimized by the MCA, $\vec{\omega}$. Unless otherwise specified, no initial filter is used and the MCA is started with unfiltered data.

The portion of the optimization equation that is dependent on the filtered data is given by

$$\sum_{i,m,t} \gamma_{t,i,m} \left[\mathcal{Y}_t - \vec{\mu}_{x,i,m}\right]^{T_r} \mathbf{\Sigma}^{-1}_{x,i,m} \left[\mathcal{Y}_t - \vec{\mu}_{x,i,m}\right] \ , \tag{3}$$

where $\gamma_{t,i,m}$ is the a posteriori probability of occupying Gaussian mixture component $m$ and state $i$ at time $t$. The quantities $\vec{\mu}_{x,i,m}$ and $\mathbf{\Sigma}^{-1}_{x,i,m}$ in Equation 3 are the HMM model means and variances for the unfiltered data, $\mathcal{X}_t$, and are not reestimated as part of this procedure. For the purposes of this development, $\mathcal{Y}_t$ in Equation 1 can be written in vector notation as:

$$\begin{aligned} \mathcal{Y}_t &= \left(\mathcal{X}_{t-b} \ldots \mathcal{X}_{t+f}\right)^{T_r} \left(\omega_{-b,i,m} \ldots \omega_{f,i,m}\right))^{T_r} \\ &= \mathcal{X}^{t+f}_{t-b}\, \vec{\omega}_{i,m} \ . \end{aligned} \tag{4}$$

Note that, while Equation 4 demonstrates that it is possible to use MS filters $\vec{\omega}_{i,m}$, that are HMM state and mixture component dependent, this is not done here. By substituting the expression for $\mathcal{Y}_t$ in Equation 3 and differentiating with respect to $\vec{\omega}$, which does not depend on $i$ and $m$, we obtain

$$\begin{aligned} \sum_{l,i,m,t} \gamma_{t,i,m} \left[\mathcal{X}^{t+f}_{t-b}\right]^{T_r} \mathbf{\Sigma}^{-1}_{x,i,m} \mathcal{X}^{t+f}_{t-b} \vec{\omega} = \\ \sum_{l,i,m,t} \gamma_{t,i,m} \left[\mathcal{X}^{t+f}_{t-b}\right]^{T_r} \mathbf{\Sigma}^{-1}_{x,i,m} \vec{\mu}_{x,i,m} \end{aligned} \tag{5}$$

Finally, $\vec{\omega}$ can be obtained by solving the matrix equation given by Equation 5 which requires the model means and variances, the a posteriori probabilities computed in the forward–backward algorithm, and the unfiltered observations.

While the above procedure can be iterated, using the estimated $\vec{\omega}$ obtained by solving Equation 5 as input to a following iteration, there a number of issues that must be dealt with. The first issue relates to the fact that it is the filtered data, as opposed to the original data, that is input to the next iteration of the algorithm. As a result, the filter applied to the data for a given iteration must be the convolution of all the filters obtained in the preceding iterations. As has been seen, a convolution is also necessary for the case where the MCA procedure is started with data filtered with an initial filter. A second issue relates to artifacts that arise due to lack of energy constraints in the model estimation procedure. An empirical procedure for normalizing filter energies between iterations is described in Section 3..

## 3. COMPENSATING FOR MS DOMAIN MISMATCH

The MCA algorithm was first applied to a simulated distortion. The goal of this first application was to implement a scenario as depicted in the block diagram in Figure 1. The form of the modulation spectrum mismatch was intended to incorporate a loose approximation to existing MS models of how speaking rate variability might be reflected in the modulation spectrum domain (e.g. [5]). These models characterize the MS of speech as having a peak at approximately four Hz with some variation in the location of that peak possibly resulting from speaking rate differences.

The modulation spectrum distortion took the form of an FIR filter of length 7 with a peak at $10Hz$ and a strong attenuation of modulation frequencies beyond $20Hz$. This filter was actually applied to the *training* utterances of the high SNR, noise–free TI digits database [4]. So the modulation spectrum shaping indicated in Figure 1 would actually be the inverse of that spectrum. The HMM model, $\lambda$, was then trained from the filtered data using the forward–backward algorithm. The 8623 digit strings uttered by a population of adult speakers in the training set of the TI digits database was used for training digit models with a mixture of at most 16 continuous Gaussian components per HMM state. The TI digits test set was split into a 3306 utterance development set, which was used as input to the MCA algorithm, and a 5376 utterance test set for evaluating ASR word accuracy.

The inputs to MCA are the model and a subset of the unfiltered development utterances and the output is a filter that reduces the mismatch between the utterances and the model. Figure 2 shows the modulation spectra for the training utterances and the unfiltered development utterances for cepstral coefficient $c_4$. The filter that was applied to the training utterances is also shown. A significant mismatch can be observed in the MS domain. The magnitude spectrum of the MS filter estimated using Equation 5 was found to provide a reasonably good approximation magnitude spectrum of the filter used on

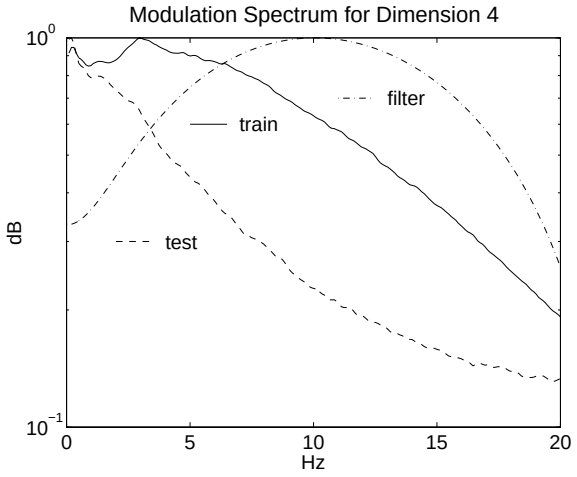



**Figure 2. Modulation Spectra for the Training Utterances and for the Unfiltered Development Utterances, along with the Spectrum of the Filter Applied to the Training Utterances**

the training utterances. However, since there is no inherent constraint on the cepstrum energy levels in the MCA, significant mismatch in the energy levels of the filtered cepstra and the original cepstra can occur. To deal with this, the filter coefficients are scaled with a dimension specific scale factor so that average energy of the training and adaptation utterances are normalized to the same level. The estimated filter, scaled separately for each component of the input cepstrum vector, is then applied to the test utterances.

Table 1 displays the recognition performance after the MS filter determined by the MCA was applied to the test utterances. Performance is presented as a function of the number of development or adaptation utterances that were used by the MCA. The algorithm was implemented here in a supervised mode with the utterance transcriptions made available to the MCA. The Table shows that when the whole development set is used to determine the best filter with which to reduce the mismatch between train and test utterances, word accuracy approaching the matched condition can be obtained. When less data is available (from 60 utterances through just one), performance degrades gracefully with respect to the whole development set and still alleviates most of the recognition rate reduction due to the MS differences between training and testing. Results are given in all cases for the filter output by the first iteration of MCA, which has length 3. Surprisingly, filters with other lengths, resulting either from running several iterations with a filter length equal to 3, or from choosing another filter length (e.g. 5, 7 or 9) also give good performance but short of the results when the length is 3. A length of 3 seems to strike the right balance between the number of degrees of freedom and a constrained estimation.

| Modulation Spectrum Compensation | Adaptation Utterances | %Word Accuracy |
|---|---|---|
| Matched | - | 96.30 |
| Mismatched | - | 15.04 |
| MCA | 3306 | 92.10 |
| MCA | 10 | 88.78 |
| MCA | 1 | 87.93 |

**Table 1. Results for MCA with a simulated distortion**

MCA seems to work well for the experiments that were done with a simulated distortion. MCA was applied to estimate the filter applied to the absolute coefficients, no dynamic features were used. The next section describes an experiment where MCA was used to optimize the filter used to obtain the delta coefficients.

## 4. MCA AND SPEAKING RATE

In order to test MCA in a more realistic environment, a database containing speech elicited at multiple speaking rates was used [6]. This database, referred to as DB1, was recorded in an anechoic room with a high quality microphone. It includes speech from 22 female, and 24 male talkers, each of which recorded 120 sentences. Both normal rate and fast rate utterances of each sentence was elicited from each speaker. Fast rate speech was elicited from each speaker by asking speakers to speak sentences as rapidly as possible without gross mispronounciations. For each speaking rate, cross-word triphones backed off to monophones were trained from the sentences from 38 speakers. The remainder were reserved for testing. Each system had 4903 states and over 28000 Gaussians. In recognition, each phone may be recognized via a triphone or the corresponding monophone at any place within a word. All 120 sentences for all the test speakers were recognized. Figure 3 describes the experiment that was carried out.

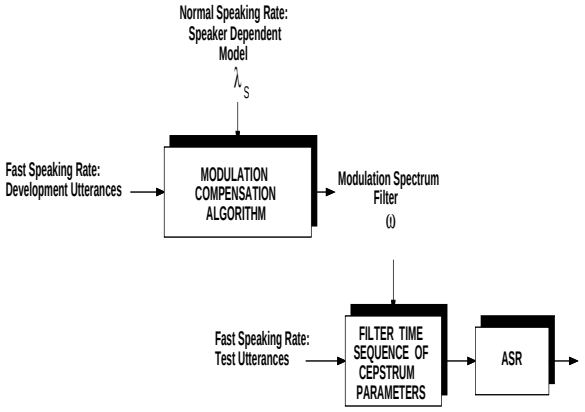



**Figure 3. MCA and speaking rate**

The goal was to study how well MCA can cope with speaking rate mismatches and up to what point these can be represented in the modulation spectrum. The recognition results for all 8 test speakers were analysed. The utterances from a single male speaker where more dramatic differences between fast and normal rate speaking styles were observed were selected for the experimental study.

Since MCA might improve word accuracy by performing speaker adaptation, a speaker adapted model was obtained. To do this, the gaussians in the normal-rate model were clustered into 4 regression classes. On the first 60 normal-rate sentences, one Maximum Likelihood Linear Regression full matrix transformation (MLLR) was learned for each regression class [3]. These were applied to the means of the normal-rate normal-rate model, which results in a speaker adapted normal-rate model.

This model was then input to the MCA along with the first 60 fast sentences, regarded as development or adaptation sentences. The filter output by MCA was applied to the 60 last fast utterances. In this case, MCA was used to optimize the filter used to obtain the delta features (no other dynamic features were used apart from them). Another difference with the previous experiment is that the filter used to obtain the delta features for the training utterances was used as the initial filter for MCA. The length

of the filter output by each iteration was the initial model filter length, i.e. 5, plus 2 for each iteration. Results for this experiment can be found in Table 2.

| Test Condition | %Word Acc. |
|---|---|
| Speaker Indep. Normal Rate | 57.28 |
| Speaker Adapt. (SA) Normal Rate | 80.67 |
| Baseline SA Fast Rate | 61.58 |
| "Pooled" MCA Fast Rate | 61.10 |
| "Dim. Specific" MCA Fast Rate | 63.96 |

**Table 2. Results for MCA with different speaking rates**

Two modes for MCA were investigated. The first, referred to as "pooled MCA" in Table 2, estimates the MCA filter coefficients according to the procedure outlined in Section 2. The second, termed "dimension specific MCA", estimates a separate set of MS filter coefficients for each dimension of the cepstrum observation vector. This represents an extension to the MS filter described in Equation 1 and independently estimates filter parameters to optimize dimension specific ML criterion. It is clear from Table 2 that the MCA yields only a modest improvement in WAC over the baseline system on the fast rate for this task.

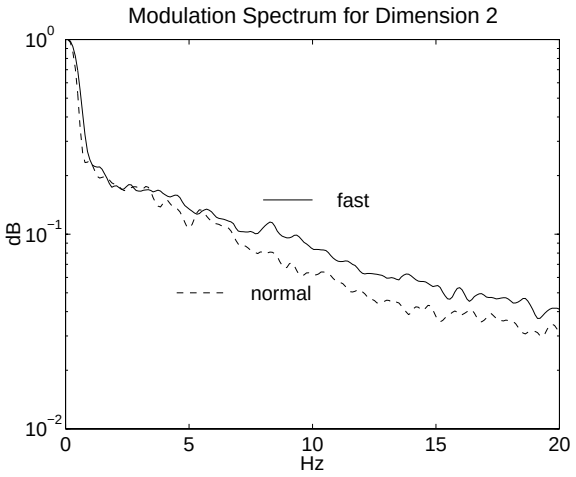



**Figure 4. Measured MS from fast and normal speech utterances**

In order to obtain some perspective for the degree to which modulation spectrum based techniques may affect performance for speaking rate mismatch, the average modulation spectrum was measured for utterances with different speaking rates. For each cepstrum component, the magnitude of the averaged modulation spectrum was computed over 60 sentences for normal and fast rate speech. Figure 4 displays the magnitude spectra for one of the components of the acoustical vector. It is clear from the curves that the fast rate speech has only slightly more energy in the higher modulation frequencies than the normal rate speech. This suggests that one might expect only small changes in performance using MS domain techniques for this task, which supports the fairly minor improvements that were observed here.

## 5. CONCLUSION

The modulation compensation algorithm was presented as a maximum likelihood technique for estimating filters for the time sequence of spectral parameters in order to reduce mismatch in the modulation spectrum. The procedure was described as an extension to a class of techniques developed in [1]. It can be applied to estimating MS filters for the absolute cepstrum coefficients or for obtaining the dynamic cepstrum coefficients from the absolute coefficients. Two applications for the MCA were described. The first application was for a task where a simulated modulation spectrum domain distortion was introduced during testing. It was shown that the MCA can effectively reduce mismatch in the MS between training and testing utterances. In a second application, the MCA was applied to reducing mismatch attributable to differences in speaking rate. The MCA was shown to have only a small impact on performance for this task. Further work is directed towards application of the algorithm to other sources of variability where distortions in the modulation spectrum are more pronounced.

## REFERENCES


[1] Rathinavelu Chengalvarayan and Li Deng. "Use of Generalized Dynamic Feature Parameters for Speech Recognition". *IEEE Trans. on Speech and Audio Processing*, 5(3):232–242, May 1997.

[2] T. Houtgast and H.J.M. Steeneken. "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria". *JASA*, 77(3):1069–1077, March 1985.

[3] C.J. Leggetter and P.C. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models". *Computer Speech and Language*, 9:171–185, 1995.

[4] R.G. Leonard. "A Database for Speaker-Independent Digit Recognition". In *Proc. ICASSP'84*, pages 42.11.1–4, 1984.

[5] Climent Nadeu, Pau Pachès-Leal, and Biing-Hwang Juang. "Filtering the time sequences of spectral parameters for speech recognition". *Speech Communication*, 22:315–332, 1997.

[6] Juergen Schroeter. "A High-Quality Speech Database". Technical report, AT&T Bell Laboratories, Murray Hill, NJ, USA, 1992.

[7] Chris J. Wellekens. "Enhanced ASR by Acoustic Feature Filtering". In *Proc. ICSLP'98*, pages 2995–2998, Sydney, Australia, 1998.