On Using Missing-Feature Theory with Cepstral Features
—Approximations to the Multivariate Integral

Frank Seide\textsuperscript{1} and Pei Zhao\textsuperscript{1,2}

\textsuperscript{1}Microsoft Research Asia, Beijing, P.R.C.
\textsuperscript{2}Department of Machine Intelligence, Peking University, Beijing, P.R.C.
\texttt{fseide@microsoft.com, pku_zhaopei@hotmail.com}

\section*{Abstract}

Missing Feature Theory (MFT), a powerful systematic framework for robust speech recognition, to date has not been optimally applied to transform-based features like MFCC or HLDAs, which are necessary for state-of-the-art recognition accuracy, due to the intractable multivariate integral in bounded marginalization. This paper seeks to enable more optimal use of MFT with MFCC features/diagonal covariances through two approximations of this integral: Numeric integration by \textit{linear sampling}, and \textit{approximation by the integral's maximum}. The former is made feasible through a “tridiagonal” approximation of MFCC, based on interpreting MFCC as bandpass-filtering the filterbank vector. The latter is solved through quadratic programming. Their effectiveness is shown for recognizing reverberated TIMIT speech utilizing temporal auditory masking.

\textbf{Index Terms:} ASR, missing features, bounded marginalization

\section{Introduction}

Missing Feature Theory (MFT)\cite{1,2,3} is rooted in recognizing “occluded” speech, where parts of the speech spectrum are obscured or damaged by other sound sources\cite{4}. Features are replaced by distributions that represent what is known about them, specifically bounded uniform distributions\cite{1}. However, this requires bounded marginalization of the multivariate Gaussian, which is intractable. Because of this, MFT could not be optimally applied to transform-based features like MFCC or HLDAs, which are necessary for state-of-the-art recognition accuracy.

Literature reports various workarounds to this conundrum: Dropping filter channels entirely (unbounded marginalization, solved by projection\cite{4,5}) or replacing them by point estimates (imputation)\cite{1,2}; integrating in MFCC space over a bounding box that circumscribes the original integration area (a trapezoid in MFCC space); or simply not using linear transforms altogether. All are suboptimal, often leading to suboptimal accuracy, limiting the widespread use of MFT. The related joint-uncertainty decoding technique (JUD)\cite{6} achieves tractability through using unbounded Gaussians instead of uniform distributions.

This paper aims to tackle this problem more directly, hoping to enable applying MFT more optimally to state-of-the-art systems. We propose two approximations to the multivariate integral: Numeric integration by \textit{linear sampling}, and \textit{approximation by the integral’s maximum}. They are evaluated on the example task of recognizing artificially reverberated TIMIT speech with MFT and a simple model of human temporal auditory masking.

\section{Missing Features; The Problem}

In this section, we will first introduce Missing Feature Theory (MFT) in the context of a broader concept of dealing with robustness, and then zoom in on the particular problem one faces when trying to apply MFT to state-of-the-art MFCC features—the untractable integral over the multivariate Gaussian density.

\subsection{Production Model with Transmission Channel}

To systematically account for distortion of the speech signal as it travels through the transmission channel (e.g. noise, filtering, reverberation, compression), we augment the generative model that underlies HMM-based recognizers with an explicit \textit{channel model}. This is depicted in Fig. 1. $W$ denotes the word sequence and $X$ the observation sequence of clean speech with language model $P(W)$ and acoustic-phonetic model $P(X|W)$. Channel model $P(Y|X)$ models how clean speech is transformed by the transmission channel into the observed distorted speech $Y$. $X$ becomes a hidden variable and is marginalized out in the decision rule for the most likely word sequence $\hat{W}$:

\begin{align*}
\hat{W} &= \arg\max_W \{P(Y|W) \cdot P(W)\} \\
P(Y|W) &= \int_X P(Y|X) \cdot P(X|W) \, dX \tag{1}
\end{align*}

For $P(X|W)$, we assume the usual continuous-density HMMs with Gaussian mixture distributions. The resulting $P(Y|W)$ is also an HMM, with mixture components of the form:\footnote{We also assume inter-frame independence of the distortion.}

\begin{align*}
p(y|m) &= \int_{R^D} p(y|x) \cdot p(x|m) \, dx \tag{2}
\end{align*}

where $y$ is an observed, distorted feature vector; $x$ the underlying clean speech vector at the same point in time; $m$ a global mixture-component index; $p(x|m)$ a Gaussian for clean speech; and $p(y|x)$ the \textit{per-frame channel model}. Both MFT and JUD\cite{6} are manifestations of this model, with different forms of $p(y|x)$. Let us now define what we mean by MFT, and then bring it into this framework.

\subsection{Missing Feature Theory In A Nutshell}

MFT shall refer to Missing Feature Theory in its pure, original sense as follows. MFT assumes that each point in the time/frequency plane is either sufficiently affected by distortion or sufficiently clean. Affected features are considered “missing data.”
MFT attempts to make use of what information is preserved in such a feature. The decision on which features are missing (the feature mask) is made by a process external to the recognizer.

MFT requires features that are separable in time and frequency. We assume our feature vectors $x$ to represent a typical front-end based on a D-channel filterbank. Features $x(t)$ shall denote the logarithmically compressed energy of filter $i$ at frame $t$.² MFCF features are not directly compatible with MFT, but can be implemented as a transform embedded in the Gaussians.

MFT further assumes that channel noise adds energy on the power spectrum of a frame, such that for missing features, the observed power spectrum value is an upper bound of the undistorted spectrum: $y_i \geq x_i$. MFT lastly assumes to know nothing further about the distribution of the distortion: The actual power spectrum of the underlying clean speech lies somewhere, with equal probability, between a meaningful lower bound $x_{\text{min},i}$ and the observed power spectrum as the upper bound: $x_{\text{min},i} \leq x_i \leq y_i$.

### 2.3. The Problem: Integrating a Multivariate Gaussian

We now express MFT w.r.t. the model of Fig. 1. MFT can be expressed by a channel model $p(y|x) = \prod_i p(y_i|x_i)$ with:

$$p(y_i|x_i) = \begin{cases} U(x_i;x_{\text{min},i},y_i) : \text{feature } i \text{ missing at } t \\ \delta(y_i - x_i) : \text{else} \end{cases} \quad (3)$$

where $U(x_i;x_{\text{min},i},y_i)$ denotes the uniform distribution over the interval $[x_{\text{min},i},y_i]$, and the Kronecker function $\delta$ acts as a “no-op” for non-missing features. For points in time where all feature dimensions are missing, Eq. (2) becomes:

$$p(y|m) = \int_{x_{\text{min}} \leq x \leq y} p(x|m) \, dx \cdot \prod_i \frac{1}{y_i - x_{\text{min},i}} \quad (4)$$

Where some components $i$ are not missing ($x_i = y_i$), the integration interval for $x_i$ degenerates to a point. Conceptually $x_{\text{min},i}$ is replaced by $y_i$ (the rather complex exact formula is not needed for this paper). If the Gaussians $p(x|m)$ have diagonal covariance matrices, the integral is trivial. Otherwise, no closed-form solution exists. The next section proposes two approximations to tackle this problem.

### 3. Approximative Bounded Marginalization

#### 3.1. The Multivariate Gaussian Integral

Eq. (4) is expressed in terms of filterbank features. To apply it to MFCF models, we factor the DCT transform into $p(x|m)$:

$$p(x|m) = \frac{1}{Z_m} \exp \left\{ -\frac{1}{2} (L M x - \mu_m)^T \Sigma_m^{-1} (L M x - \mu_m) \right\} \quad (5)$$

where $\mu_m$ and $\Sigma_m$ are mean and (diagonal) covariance matrix of mixture component $m$ of an underlying clean-speech model trained on MFCF features. $L M$ accounts for the MFCF transform: $M$ is the DCT ($M M^T = I$), and the diagonal $L$ accounts for any scaling like liftering. $Z_m$ normalizes the density.

However, with this, Eq. 4 has no closed-form solution. This is the reason why MFT cannot be easily combined with linear-transform based features like MFCF or HLDA.

We introduce two approximations to tackle the problem: Numerical integration through uniform sampling (in combination with an approximation of the MFCF transform); and approximating the integral by the maximum.

²In MFT literature, a popular alternative are features that more closely resemble the human auditory system, e.g. [7].

³This implies the distortion to be independent between filterbank channels. This reflects standard MFT, but we suspect this is not correct.

### 3.2. Numeric Integration; CDFF Approximation of MFCF

The usual Monte-Carlo approach to this integral [9] would sample the space inconsistently across Gaussians, risking problems in the recognition process. Instead, we approximate the integral by nested sums over $K$ equidistant points spanning $[x_{\text{min},i}, y_i]$:

$$p(y|m) \approx \frac{1}{K^D} \sum_{x_1} \cdots \sum_{x_D} p(x|m) \quad (6)$$

To make this feasible, we further approximate MFCF through a tridiagonal transform, as follows. Eq. 5 can be written as:

$$p(x|m) = \frac{1}{Z_m} \exp \left\{ -\frac{1}{2} (H_m x - \mu'_m)^T (H_m x - \mu'_m) \right\} \quad (7)$$

with $H_m = M^T \Sigma_m^{-0.5} L M$ and $\mu'_m = M^T \Sigma_m^{-0.5} \mu_m$. $H_m$ has the form of a filtering operation applied to $x$, with the diagonal $(\Sigma_m^{-0.5} L)$ as the “transfer function” [8]. This is because the DCT is a special case of the Fourier transform for real-valued symmetrical periodic signals. Figure 2 shows what $x$ gets convolved with for the case of a globally pooled covariance matrix (identical $H_m$ across all Gaussians)—a bandpass filter that essentially subtracts from each filterbank channel about a third of its direct neighbors.

i.e., MFCF combined with variance weighting constitutes a bandpass filter on the symmetrically and periodically extended filterbank-coefficient vectors³ (note that $M^T$ cancels out).

We can now make a tridiagonal approximation of MFCF by cutting down this bandpass to three components (for the example of Fig. 2, we get $h_m \propto (-0.34, 1.0, -0.34)$). With a tridiagonal $H_m$, Eq. 6 can be evaluated using a process similar to forward expansion, reducing evaluation from $O(K^D)$ to $O(K^3 D)$.

In our experiments, this approximation, which we call Cepstrally Decorrelated Filterbank Features (CDFF), costs about 1% point of accuracy, which is sufficiently close for the purpose of this paper. While still too costly for realtime operation, it brings computation into feasible range for the purpose of studying missing-feature recognition, which was our goal.

### 3.3. Maximum Approximation

A more aggressive but ultimately more useful approximation is to approximate the integral by the maximum of the integrand, based on past experience that for many speech-recognition algorithms, sums of probabilities can be approximated by their maximum at little loss of accuracy:

$$p(y|m) = C \int_{x_{\text{min}} \leq x \leq y} p(x|m) \, dx \approx \max_{x_{\text{min}} \leq x \leq y} \left\{ p(x|m) \right\} \quad (8)$$

If we likewise replace the Gaussian-mixture sums by maxima, then a “closed-form” solution exists, boiling down to a bounded minimization that can be solved by quadratic programming:

$$\begin{align*}
\text{minimize} & \quad (x - \nu_m)^T P_m (x - \nu_m) \\
\text{subject to} & \quad x_{\text{min},i} \leq x_i \leq y_i \quad \text{if } x_i \text{ “missing”} \\
& \quad x_{\text{min},i} = y_i \quad \text{otherwise}
\end{align*} \quad (9)$$

with $P_m = H_m^T H_m$ and $\nu_m = (H_m^T)^{-1} \mu_m$. Compared to linear sampling, it does not incur a loss in accuracy according to our experiments; is computationally cheaper; and has the strong advantage that it can be used with the original MFCF without the tridiagonal CDFF approximation. Actually, it is admissible to any
linear transform including HLDA, MLLR, and full-covariance models (although for those we have so far no experimental data to validate its effectiveness w.r.t. recognition accuracy).

4. Test Case: Reverberated Speech

The underlying “big-picture” motivation for this work has been to study how human auditory masking may affect automatic speech recognition. Auditory masking is a process by which signal components in the time-frequency plane can make other, weaker signal components in their time-frequency neighborhood imperceptible. Our hypothesis is that auditory masking is an important factor in the robustness of human speech recognition, and to study this, we need to teach recognizers to do as humans do—ignore masked information—using the mathematical tool of MFT.

One example is reverberation and temporal auditory masking [3]. Reverberation in face-to-face conversations in daily-life rooms typically fades faster than the human temporal mask, suggesting imperceptibility of most reverberation. We chose this as the test case for our approximations: Temporal auditory masking yields a feature mask for use in MFT-based recognition (Eq. 3).

For illustration, Figure 3 plots the log filterbank channel energy for both clean speech and in three reverberant conditions. The exponentially decaying reverberation tails distort the lower-energy regions. High-energy portions remain largely unaffected.

Our model for temporal masking is a crude exponential decay: \( L_i(t) = \max\{L_i(t - \Delta T) - \Delta T/\tau; y_i(t)\} \), with masking threshold \( L_i(t) \) (log domain like our features), frame shift \( \Delta T \), and \( \tau \) matching the human’s auditory mask decay (\( T_\infty \approx 350\text{ms} \)). The resulting masking curves are also shown in Fig. 3 (the straight gray downward lines topping off the “valleys”).

As we will see later, when recognizing reverberated speech with a clean-speech model, this simplistic approach—which does not even use any estimate of reverberation—recovers a significant part of the accuracy loss due to reverberation. However, we reemphasize that this method, despite being novel, is not the point of this paper. Used by itself, its effectiveness is still limited and not competitive (one would involve multiple microphones, multi-condition adaptive training, model adaptation, and a more refined online-adaptive channel model beyond MFT). For this paper, its sole purpose is to validate our proposed approximations.

5. Results

The approximations were evaluated on the TIMIT corpus, using the standard training/test split (about 3.1 hours of training data; 192 test utterances), 3740 cross-word triphones with CART tying to 1003 states, a phoneme bigram trained on the training transcripts, and the usual TIMIT phoneme mapping. All models were trained using HTK on clean speech only. The “vanilla” baseline system achieves the usual phone-error rate reported for TIMIT of around 33%. Except for a slightly reduced frequency range to 70–7000 Hz, filterbank and MFCC features use HTK’s default settings unless otherwise noted.

Three reverberated versions of the test set were created artificially by convolving the files with room impulse responses we measured for several rooms at various microphone distances. Impulse responses were grouped into three groups with \( T_\infty \) of 100–150, 150–200, and 200–250 ms, respectively. Recognition was done with wide pruning (400). Language-model weights were roughly hand-tuned on the clean set to balance insertions and deletions.

For all MFT experiments, we modified Cepstral Mean Normalization (CMN) to derive the utterance mean from the masking curve \( L(t) \) rather than \( y(t) \) (training+test). This makes it mostly invariant to the large variation in the “valleys” due to reverberation. This improves accuracy for reverberated speech by about 1% point while hurting less than 0.5% points for clean speech.

Since MFT only provides tight bounds for the static features, we also include results without delta or acceleration coefficients.

5.1. Baselines; Effect of CDFF Approximation of MFCC

Table 1 shows results for various baselines, starting with the “vanilla” system (phone-error rate of 32.7%).

Because of the initially uncertain outcome of this investigation, we took two implementational shortcuts: We used a square DCT transform (i.e. using only \( D = 13 \) filter channels instead of HTK’s default of 20) and a pooled covariance matrix. The table validates that the incurred degradation is below about 3% points.

Neither shortcut is required for our approximations, but they greatly simplified the CDFF model training: With a pooled covariance matrix, CDFF’s filter \( h_m \) is independent of the individual Gaussian \( m \), and can thus be implemented for model training as an external feature transform. The \( h_m \) was computed from the variances of the static features of a corresponding MFCC model, and fixed during training. The same \( h_m \) was also used for deltas and accelerations. Table 1 also shows the baseline without deltas and accelerations, which is 8–12% points worse.

5.2. Numeric Integration Under CDFF Approximation

Table 2 shows results for clean and reverberant speech. To separate effects, this comparison is shown for static-only features.
The ‘base’ column uses no missing features—showing the effect of reverberation if nothing is done about it. The column headed \( \int \) uses precise integration for Eq. 4. This is only possible for FBANK features. 43–49% of all features were “missing.” The column headed \( \sum \) shows results for linear sampling of \( K = 10 \) points per Eq. 6. The row “Av. loss reduction” shows how much of the error-rate increase is recovered through the respective MFT setup, averaged over the three reverberation conditions. It indicates the effectiveness of the respective MFT approach. First, for FBANK, linear sampling is no less effective than precise integration (slightly better, 37% instead of 33% loss reduction). Secondly, the same effectiveness is achieved for CDFF. We take this as evidence that MFT with bounded marginalization is applicable and effective with MFCC (of which CDFF is a close approximation) and that our approximation indeed enables that.

The table also shows that using MFT on clean speech degrades accuracy to about the same accuracy as for 100-150 ms reverberation. I.e., the recognizer does benefit from the supposedly imperceptible low-energy speech in the “valleys.”

### 5.3. Maximum Approximation

Table 3 shows results for maximum approximation (column label ‘max’). The FBANK and CDFF results are very similar to the linear-sampling results in Table 2. The maximum approximation also works for unapproximated MFCC: As for clean speech, MFCC results are in the order of 1% point better than CDFF.

We repeated the previous experiment with full features including deltas and accelerations (Table 4). For those, the precise channel model \( p(y|x) \) would have the form of a convolution of the channel models of the underlying static features, but for simplicity, we again assumed a uniform distribution. Its bounds are based on the combining the intervals of the involved static features. About 75% of deltas and 90% of accelerations were affected. Due to the weaker models for deltas and accelerations, we expected a lower effectiveness. Indeed, Table 4 shows weaker loss reductions of 26 and 28% for CDFF and MFCC, respectively.

### 5.4. Discussion

We believe that the above results allow us to reasonably conclude, albeit indirectly, that for MFCC features, the maximum approximation of the multivariate integral is a good approximation w.r.t. the recognition results that would be achievable if the integral was tractable for MFCC.

We believe so because we have shown (for FBANK) that approximation by linear sampling is as accurate as precise integration, and we see no reason why this would be different for CDFF. Next, we find the maximum approximation also to have the same accuracy for CDFF. And since CDFF is a close approximation of real MFCC (≤ 1% point difference across nearly all directly comparable results), it is reasonable to assume what is true for CDFF is also true for real MFCC.

Amongst the two approximations, the maximum approximation would be the better choice for most practical purposes, because it is faster and, more importantly, works for real MFCC features (is not restricted to the tridiagonal CDFF approximation), and is also admissible for any other linear transform or covariance structure. Linear sampling, on the other hand, would allow to experiment with arbitrarily-shaped bounded channel models as needed, e.g., if we were to implement a more precise convolution-based channel model of delta features.

### 6. Conclusions

Aiming at being able to optimally apply MFT to MFCC features, which requires the intractable bounded integration of the multivariate Gaussian, we have shown that a workable solution is to approximate the integral by its integrand’s maximum within the integration region, which is tractable through quadratic programming. Taken together, our results on TIMIT data reasonably suggest that it yields similar accuracies for MFCC to what one would get with the true integral if it were tractable.

The maximum approximation is also admissible to any other linear transform (e.g. HLDA and MLLR) or covariance structures, although the effectiveness for other types of transform remains to be experimentally verified. The approximations were tested for recognizing reverberated speech, where mimicking human auditory masking yielded feature masks for subsequent missing-feature recognition. As a side result, we saw that MFCC can be interpreted as decorrelating filterbank channels by “band-pass filtering” the feature vector, a physiologically plausible interpretation.

We hope to eventually enable research into the more direct model-based application of missing-feature theory on state-of-the-art baselines. This paper is a step towards this.

### 7. Acknowledgements

We thank Lie Lu, Alfred Mertins, Mike Seltzer, and Kit Thambiratnam for their support, ideas, and valuable feedback. The room impulses were measured by Minhua Chen during an internship.

### 8. References


### Table 3: Maximum approximation of integral. Static features only.

<table>
<thead>
<tr>
<th>Reverberation</th>
<th>FBANK base max</th>
<th>CDFF base max</th>
<th>MFCC base max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>51.3 54.4 47.8 30.6 46.0 48.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100-150 ms</td>
<td>57.5 54.5 54.9 31.9 55.5 51.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>150-200 ms</td>
<td>64.1 58.6 62.5 57.2 61.9 55.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200-250 ms</td>
<td>68.4 63.4 66.6 61.3 66.0 60.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Av. loss reduction</td>
<td>40% -36% -39%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Maximum approximation of integral. Full features.

<table>
<thead>
<tr>
<th>Reverberation</th>
<th>FBANK base max</th>
<th>CDFF base max</th>
<th>MFCC base max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>43.9 49.6 35.5 42.2 34.3 40.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100-150 ms</td>
<td>52.4 49.7 44.8 43.8 44.2 42.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>150-200 ms</td>
<td>63.0 54.1 55.6 48.6 55.1 48.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200-250 ms</td>
<td>68.3 60.8 63.7 48.8 62.4 54.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Av. loss reduction</td>
<td>37% -26% -28%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>