ITERATIVE GROUP SELECTION-BASED ENHANCEMENT OF TIME-FREQUENCY MASKS FOR MISSING DATA RECOGNITION

DANIEL PULLELLA* and ROBERTO TOGNERI†

School of Electrical, Electronic and Computer Engineering
The University of Western Australia, Crawley WA 6009, Australia
*daniel.pullella@gmail.com
†roberto.tongneri@uwa.edu.au

Received 31 May 2011
Accepted 25 May 2012
Published 3 September 2012

Missing data approaches have recently been applied to speech recognition tasks to increase noise robustness. The drawback of missing data techniques is the vulnerability of the recognizer to errors in the reliability mask. This work proposes a novel group selection algorithm to perform top-down refinement of initial bottom-up reliability mask estimates with the goal of removing these errors. A novel probabilistic decision process based on normalized likelihood distances is proposed and used to evaluate the quality of a reliability mask without any a priori noise knowledge. Experimental results on a speaker identification task illustrate the ability of the combined bottom-up top-down system to significantly outperform traditional bottom-up only missing data techniques for various types of mask corruption.

Keywords: Missing feature theory; missing data recognition; time-frequency masking; iterative group selection.

1. Introduction

In speech recognition there have been sophisticated advances on the standard state-of-the-art features and models dealing with various forms of robustness issues, from system improvements when dealing with non-neutral speech to efficient measures of confidence of the decoding process. In this paper we adopt a similar sophisticated approach to robust speaker recognition in the presence of noise. We propose a novel machine learning solution to improve the identification rate through use of the missing data paradigm. Although presented in the context of speaker identification the proposed machine learning solution is general enough to be adopted in any application requiring reliable mask estimates with minimal prior knowledge.

† Corresponding author.
Missing data approaches\(^\text{9,11,34,36}\) have been shown to be effective in compensating against arbitrary disturbances within a speech signal and are thus naturally suited to dealing with the problem of environmental distortion. Conceptually, these techniques are based on the quantification of the noise in each individual time-frequency (TF) region of the input speech signal. At a given TF point a low signal-to-noise ratio (SNR) indicates noise dominance and hence an ‘unreliable’ component, while a high SNR indicates speech dominance and hence a ‘reliable’ component. The assignment of reliability decisions to all TF points within the speech signal allows the construction of a TF reliability mask which enables missing data compensation to be performed either through the reconstruction of the unreliable spectral components,\(^\text{33}\) or through the modification of the recognizer to utilize marginal probabilities based on the reliable components.\(^\text{11}\)

Regardless of the specific recognition paradigm used, it is clear that the robustness provided by missing data compensation is critically dependent on the accuracy with which the reliability decisions are made. If a\(a\)\(\text{priori}\) noise knowledge is available reliability labels can be assigned with perfect accuracy based on the true SNR at each TF point. The resulting oracle masks allow missing data methods to provide high robustness even in conditions of extreme noise corruption.\(^\text{10}\) However, a\(a\)\(\text{priori}\) noise knowledge is not always available, and in practice other techniques are required to estimate the oracle mask. The use of noise estimation techniques such as spectral subtraction\(^6\) to find the local SNR at each TF point provides an intuitive approach to estimating the oracle reliability mask. Using an appropriate SNR threshold for the reliability decision, spectral subtraction-based missing data has demonstrated significant performance improvements for speech and speaker recognition in stationary disturbances.\(^\text{14,22,44}\) The drawback of spectral subtraction estimation is its lack of robustness to nonstationary noises.

Previous research has also utilized computational auditory scene analysis (CASA) to perform missing data mask estimation. Based on the desire to identify, separate and process sounds from individual sources, these approaches can be applied to segregate regions of the TF domain that are target speaker dominated from regions which are dominated by some other source (such as background noise, interfering speakers, etc.).\(^\text{4,8,17}\) Specifically pitch-based monaural voiced speech segregation\(^\text{18}\) has been successfully used to perform missing data mask estimation in robust speaker identification.\(^\text{39}\) Voicing information has also been used in combination with traditional reliability decisions to enhance speaker identification robustness.\(^\text{19}\) Based on a spectral shape algorithm to identify regions of voiced speech\(^\text{20}\) a voicing reliability mask can be constructed, and its combination with a traditional local SNR-based reliability mask provides improvements over the use of the SNR mask alone.

In addition to bottom-up methods which utilize the deterministic properties of the signal, band specific model training has also been used to perform TF reliability mask estimation. Under a statistical Bayesian approach TF points are tagged as reliable or unreliable according to the distributions of a set of feature parameters extracted from
the speech observations. Such estimation has been successfully implemented for Gaussian mixture classification using spectral energy with derivatives, and combined energy and statistical features representing speech characteristics. This latter work is extended by incorporating spectral variation across time and frequency to improve upon the simple white noise training initially proposed. Alternatively hard, binary masks can be replaced by soft masks utilizing likelihood or fuzzy decisions from probabilistic or cluster modeling of the feature space. Recent research has also proposed the use of relevance vector machine (RVM) classification to construct the reliability mask through the direct modeling of the Short-Time Fourier Transform (STFT) coefficients. The results suggest improved performance compared to CASA pitch-based estimation, but at a greatly increased computational cost.

The weakness of the traditional approaches to missing data mask estimation is that the produced masks often contain errors: the inclusion of truly unreliable points and the deletion of truly reliable points. Since the missing data recognizer has no protection against these errors, recognition rates obtained from estimated masks are typically far below those obtained using ideal reliability masking. To solve this problem recent research has attempted to increase the accuracy of the reliability decisions by combining bottom-up estimation with top-down methods which utilize the trained models. For speech recognition examples include multi-source decoding which performs a simultaneous search in the fragment-labeling space and the word string space to find the most likely overall solution, the two-stage speech separation hypothesis testing approach which integrates mask estimation with the recognition process, using non-negative matrix factorization to discover phone-sized time-frequency patches from which missing data masks are estimated, and the idea of Bayesian mask estimation based on statistical models of the reliable and unreliable features. One of the advantages of the application of missing data recognition techniques to speech recognition is that sophisticated mask estimation techniques exploiting phonetic/linguistic specific TF structure can be utilized.

Top-down processing is utilized for speaker recognition in the universal compensation technique which combines multi-condition training and missing data theory to compensate against arbitrary corruption types. Using probabilistic union modeling, a search is performed over the feature space of each multi-condition model to find components matched to the input spectrum allowing the formation of model specific feature subsets which maximize recognition performance. The disadvantage of this approach is the required exhaustive search over the feature space of each of the model spectra, since this may be computationally intensive for high feature dimensions. Unlike speech recognition applications, speaker recognition cannot directly utilize source specific language knowledge without access to a speech decoder (e.g. HMM system).

In this paper we propose the use of top-down processing to refine initial bottom-up missing data mask estimates applied to the task of robust speaker identification. Given a binary estimated reliability mask, the goal of the top-down stage is to
remove mask errors and reproduce the oracle mask. The use of the oracle mask as the objective for the top-down processing allows the proposed system to perform recognition efficiently by avoiding the calculation of model specific feature subsets, while still providing high robustness. In a continuation of previous work the normalized segment likelihood distance (NSLD) criterion is utilized within the top-down refinement process to evaluate the quality of candidate TF reliability masks. This criterion is based on the model likelihood distributions produced by accumulated marginal densities, and can discriminate between oracle and inaccurately estimated TF masks without a priori noise knowledge. A novel group selection algorithm is proposed to perform the mask refinement. Using sparsity and frequency reliability criteria, refinement proceeds iteratively via the weighted selection of a group of TF points which are potential inclusion errors, and the subsequent removal of this group from the mask. The normalized likelihood distance is used to calculate the quality of the resulting mask, and through the use of probabilistic replacement decisions this ensures that refinement proceeds towards the oracle mask when initialized with an inaccurate estimate. An experimental evaluation of the proposed system was performed on the TIMIT database for varying strengths of additive noise distortion. The results demonstrated an improvement when using the combined bottom-up estimation top-down enhancement system compared to both standard bottom-up only missing data, and a system which uses simple median filtering-based mask post-processing.

The remainder of this paper is organized as follows. Section 2 provides an overview of GMM speaker identification with missing data marginalization. The normalized likelihood distance criterion is reviewed in Sec. 3, followed by Sec. 4 which presents the iterative group selection algorithm and the combined bottom-up top-down identification system. An experimental evaluation of the proposed system is presented in Sec. 5. Conclusions and avenues of future work are presented in Sec. 6.

2. Speaker Identification with Missing Data

2.1. Gaussian mixture model classification

In Gaussian Mixture Model (GMM) speaker identification each speaker is represented as a weighted sum of M Gaussian densities. Given a speaker represented by model λ, and a D-dimensional spectral observation vector \( \mathbf{x} = (x_1, x_2, \ldots, x_D)' \), the observation likelihood is

\[
p(\mathbf{x} | \lambda) = \sum_{i=1}^{M} c_i \mathcal{N}(\mathbf{x}; \mu_i, \Sigma_i),
\]

where \( c_i \) is the weight of the \( i \)-th mixture and \( \mathcal{N} \) is a \( D \)-variate Gaussian

\[
\mathcal{N}(\mathbf{x}; \cdot) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} e^{-\frac{1}{2}(\mathbf{x} - \mu_i)'(\Sigma_i)^{-1}(\mathbf{x} - \mu_i)}
\]
Iterative Group Selection-Based Enhancement of TF Masks

with mean vector $\mu_i \in \mathbb{R}^D$ and covariance matrix $\Sigma_i \in \mathbb{R}^{D \times D}$. Each speaker is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities:

$$\lambda = \{c_i, \mu_i, \Sigma_i\}, \quad i = 1, 2, \ldots, M. \quad (3)$$

2.2. Marginalization recognition

Let the binary TF mask vector corresponding to observation $x$ be $m = (m_1, m_2, \ldots, m_D)'$. Based on this mask vector the components of a given observation can be labeled as reliable ($r$) or unreliable ($u$) according to the distribution of the unreliable components:

$$x = (x_r, x_u)', \quad \mu_i = (\mu_{ri}, \mu_{ui})' \quad \text{and} \quad \Sigma_i = \begin{bmatrix} \Sigma_{rr_i} & \Sigma_{ru_i} \\ \Sigma_{ur_i} & \Sigma_{uu_i} \end{bmatrix}. \quad (4)$$

The marginal probability density $p(x_r | \lambda)$ is obtained by integrating over the distribution of the unreliable components:

$$p(x_r \mid \lambda) = \sum_{i=1}^{M} c_i \mathcal{N}(x_r; \mu_{ri}, \Sigma_{rr_i}) \int_{x_L}^{x_H} \mathcal{N}(x_u; \mu_{ui}, \Sigma_{uu_i}) dx_u, \quad (5)$$

where $\mu_{r|ri} \in \mathbb{R}^{D_r}$ and $\Sigma_{r|rr_i} \in \mathbb{R}^{D_r \times D_r}$ are the conditional mean and conditional covariance respectively. In bounded marginalization recognition the lower and upper integration bounds are set to 0 and the observed component value respectively such that $[x_L, x_H] = [0, x_u]$, and marginal density thus becomes

$$p(x_r \mid \lambda) = \sum_{i=1}^{M} c_i \mathcal{N}(x_r; \mu_{ri}, \Sigma_{rr_i}) \int_{0}^{x_u} \mathcal{N}(x_u; \mu_{ui}, \Sigma_{uu_i}) dx_u. \quad (6)$$

For a group of $S$ speakers with corresponding GMMs $\lambda_s$, $s \in S = \{1, 2, \ldots, S\}$ the identification decision for an observation $x$ is found by maximizing the posterior probability over all models:

$$\hat{s} = \arg \max_{1 \leq s \leq S} p(\lambda_s \mid x) = \arg \max_{1 \leq s \leq S} \frac{p(x \mid \lambda_s) p(\lambda_s)}{p(x)}. \quad (7)$$

For marginalization the observation likelihood $p(x \mid \lambda_s)$ must be replaced by the marginal density defined in (5). Ignoring the constant normalization probability $p(x)$ and assuming equiprobable speakers, the decision for an utterance $X = (x_1, x_2, \ldots, x_T)$ is obtained by maximizing the log-likelihoods over all observations:

$$\hat{s} = \arg \max_{1 \leq s \leq S} \sum_{t=1}^{T} \log p(x_t \mid \lambda_s). \quad (8)$$
3. Likelihood-Based Mask Quality Criterion

The ability to successfully refine a TF mask is dependent on the availability of a measure which can discriminate between the oracle mask and inaccurately estimated or erroneous masks. Recently the normalized likelihood distance has been proposed as a criterion which is able to make such a distinction without \textit{a priori} noise knowledge.\(^{31}\) This section provides a review of the normalized segment likelihood distance criterion, which relates the accuracy of the reliability mask segment to its likelihood score confidence over all speaker models.

3.1. Mask likelihood definitions

In the following a parameterized speech utterance of length \(T\) frames is represented by a sequence of observation segments, each of size \(K\) frames. Consider an observation segment \(X = \{x_1, x_2, \ldots, x_K\}\), with a corresponding reliability mask segment \(M = \{m_1, m_2, \ldots, m_K\}\) such that \(M(t, f) = m_{tf}\) is the reliability decision for \(x_{tf}\), component \(x_f\) in observation \(x_t\). The likelihood score of the segment \(X\) on a given model \(\lambda\) is computed as the sum of the marginal likelihoods over all observations:

\[
L_\lambda(X | M) = \sum_{k=1}^{K} \log p(x_{k} | \lambda),
\]

where \(p(x_{k} | \lambda)\) is defined as in (5), and the reliability partitioning is determined by \(m_k\). The distribution of segment likelihoods for all speaker models is defined as

\[
D(X | M) = \bigcup_{\lambda \in S} \{L_\lambda(X | M)\}. \tag{10}
\]

For a diverse speaker population we define two distinct speaker sets based on the mask specific likelihood distribution \(D(X | M)\): the speaker \(s_{\text{max}}\) scoring with the maximal distribution likelihood value

\[
L_{\text{max}}(X | M) = \max_{s \in S} L_\lambda(X | M), \tag{11}
\]

and the set of competitor speaker models defined as the subset of \(N\) speakers with the highest accumulated likelihood scores, excluding \(s_{\text{max}}\). The distribution of competitor segment likelihoods is given by

\[
D_{\text{comp}}(X | M) = \{L_\lambda(X | M) | s \in Z\}, \tag{12}
\]

where \(Z \subseteq S\) with \(|Z| = N\) such that

\[
Z = \arg \max_{|V|=N, V \subseteq S, s_{\text{max}} \notin V} Q(V), \tag{13}
\]

and \(Q(\cdot)\) is the sum of the segment likelihoods from all models whose speakers are included in the set:

\[
Q(V) = \sum_{s \in V} L_\lambda(X | M). \tag{14}
\]
By applying the arithmetic mean operator $E[\cdot]$ to the competitor likelihood distribution $\mathcal{D}_{\text{comp}}(\mathcal{X} | \mathcal{M})$, a single likelihood value $E[\mathcal{D}_{\text{comp}}(\mathcal{X} | \mathcal{M})]$ is produced to represent the competitor model set.

### 3.2. Oracle mask properties

Consider the oracle mask segment where reliability decisions are assigned utilizing 

* a priori

noise knowledge. Since reconstruction of the oracle mask is the goal of the mask refinement process, it is assumed that the oracle segments produce a likelihood distribution where the maximal likelihood value $L_{\text{max}}$ is produced by the speaker model to which the observation segment truly belongs. The set of competitor models producing likelihoods in $\mathcal{D}_{\text{comp}}$ therefore correspond to the highest scoring imposter speakers.

In the case of the oracle mask the reliable component contribution of each observation $x_r$ contains only truly reliable points, and as a result the true model likelihood score is typically large. Imposter models will receive significantly lower accumulated likelihood scores due to mismatch between their trained distributions and the observed reliable components, and also due to low likelihoods from the integral contribution (in the case where the distribution means exceed the observed unreliable component values). This conceptual analysis suggests that the accumulation of bounded marginal likelihoods produced by the oracle mask will strongly favor the true model over imposter (competitor) models.

### 3.3. Normalized likelihood distance error detection

Consider a binary estimated reliability mask which contains inclusion and deletion error corruption. Compared to the oracle mask, estimated masks with either type of error experience a decrease in the likelihood score of the true model, and an increase in the likelihood score obtained by one or more imposter models. We define inclusion errors when $m_{tf} = 1$, but $x_{tf}$ is in fact unreliable or noisy (noisy speech TF points are incorrectly included) and deletion errors when $m_{tf} = 0$, but $x_{tf}$ is in fact reliable or noise-free (clean speech TF points are incorrectly excluded).

For inclusion errors (see Fig. 1(a)) the true model likelihood is decreased due to the mismatch between the trained distributions and the observed values of the inclusion error components, while imposters may benefit from a decrease in the likelihood penalty (since some truly noisy components no longer contribute to the multi-variate integral). For deletion errors (see Fig. 1(b)) there is a decrease in the true model likelihood, since observed components appear in the integral contribution increasing the counter-evidence. For imposter models with low distribution means, the deleted components receive an increase in their accumulated likelihood score due to the removal of the reliable component mismatch at the cost of a negligible marginalization penalty.

Based on this analysis, a likelihood confidence hypothesis is developed to relate the distribution of accumulated likelihood scores which are produced by an estimated
Fig. 1. The effect of unreliable inclusion (a) and reliable deletion (b) errors on model likelihood scores using binary masking with bounded marginalization. Inclusions primarily cause a decrease in the likelihood of the true model, while deletions increase the likelihood of imposter models. Single mixture diagonal covariance modeling is shown to demonstrate the effect of incorrect reliability decisions for an individual spectral component.
mask segment to the accuracy of its reliability decisions compared to the oracle mask segment. This hypothesis is summarized as follows:

**Hypothesis 1.** The oracle mask produces both a large maximal likelihood score, and a large difference between this maximal likelihood and likelihoods of the nearest competitor models (given a sufficient number of observations). The correction of all errors within an estimated mask should thus increase the maximal likelihood value (\(L_{\text{max}}\)), and increase the maximal-to-competitor model likelihood distance (\(L_{\text{max}} - E[D_{\text{comp}}]\)).

Based on this hypothesis the normalized segment likelihood distance (NSLD) is defined to quantify the mask likelihood confidence, and consequently the quality of segment \(\mathcal{M}\):

\[
\text{NSLD}(\mathcal{X} \mid \mathcal{M}) = \frac{L_{\text{max}}(\mathcal{X} \mid \mathcal{M}) - E[D_{\text{comp}}(\mathcal{X} \mid \mathcal{M})]}{L_{\text{max}}(\mathcal{X} \mid \mathcal{M})}. \tag{15}
\]

Evaluations of the NSLD in previous work have demonstrated the validity of the mask likelihood confidence hypothesis for binary error detection in both artificially distorted and practically estimated masks, where it was shown that significantly lower NSLD values were observed compared to the true oracle mask over a range of SNR values in white noise.

4. **Iterative Group Selection Refinement**

This section proposes a system to enhance the robustness of missing data speaker recognition via the enhancement of estimated TF masks using iterative group selection. Given an initial TF reliability mask estimate, the refinement of each mask segment proceeds by the iterative selection of TF points which are potentially errors where the probability of selection is determined by sparsity and energy criteria. Using the normalized likelihood distance (Sec. 3) as a measure of mask quality, the removal of groups of TF points is performed probabilistically resulting in the elimination of inclusion errors from the initial bottom-up mask (see Fig. 2).

4.1. **Initial TF mask estimation**

The group selection algorithm requires an initial estimated TF reliability mask on which refinement can take place. If a priori noise knowledge is available then the ideal oracle mask can be constructed according to the 0 dB criterion:

\[
m_{\text{oracle}}^{\text{if}} = \begin{cases} 1 & \text{if } x_{\text{if}}^s > x_{\text{if}}^n, \\ 0 & \text{otherwise}, \end{cases} \tag{16}
\]

where \(x_{\text{if}}^s\) and \(x_{\text{if}}^n\) are the clean speech and noise spectral energies, respectively. Since a priori noise knowledge is not available in practice spectral subtraction masking is also considered in this implementation. For estimated clean speech and noise spectral energies, components of the binary spectral subtraction mask are given
Fig. 2. The iterative group selection speaker identification system. Bottom-up estimation provides an initial reliability mask which is refined using top-down enhancement based on frame likelihoods to remove errors.
by:

\[ m_{if}^{ss(\theta)} = \begin{cases} 
1 & \text{if } 10\log_{10}\left(\frac{x_{sf}}{x_{nf}}\right) > \theta, \\
0 & \text{otherwise},
\end{cases} \tag{17} \]

where \( \theta \) is the energy threshold in dB.

### 4.2. Group selection using sparsity and energy criteria

Consider an observation segment \( X \) with corresponding binary mask segment \( M \) as defined in Sec. 3. The first step of the refinement process is the selection of a set of candidate inclusion error points which are to be removed from the segment. Only inclusion errors within the binary estimated mask segment are considered, and this is motivated by the desire to constrain the search space traversed by the algorithm while still removing the most severe mask errors.

Selection of a group of \( G \) points is performed via stochastic random sampling (SRS) with no replacement (the roulette method). Although other selection techniques such as stochastic universal sampling (SUS) have simpler time complexity, the computation required for group selection is negligible compared to objective function evaluation and so SRS is chosen due to simplicity of implementation.\(^1\) The probability of selection for each individual point within the mask is determined by the use of two criteria. Sparsity is the primary selection measure, which is based on the property that included TF points in isolated regions of the mask space are more likely to be truly unreliable compared to an included point within a cluster. The sparsity of each TF point in the mask segment is calculated by considering the mask value of neighboring points, each of whose sparsity weight contribution is determined by a neighborhood map \( B \). In this implementation a \( 5 \times 5 \) diamond shaped map is used with points directly adjacent to the reference point having a higher contribution than points which are further away (see Fig. 3).

Sparsity weights are calculated as the reciprocal of the contribution of the neighboring points divided by the total number of points which contribute:

\[ w_{sp}(t, f) = \mathcal{M}(t, f) \times \left[ \sum_{n=n_L}^{n_H} \sum_{q=q_L}^{q_H} \frac{\mathcal{M}(n, q) \times B(k, l)}{(n_H - n_L + 1) \times (q_H - q_L + 1)} \right]^{-1}, \tag{18} \]

where \( n_L = \max(1, t - 2), \ n_H = \min(K, t + 2), \ q_L = \max(1, f - 2) \) and \( q_H = \min(D, f + 2) \) define the ranges of the time and frequency indices \( n \) and \( q \) for the neighborhood points within the mask segment, and \( k = n - (t - 2) + 1 \) and \( l = q - (f - 2) + 1 \) are the relative indices used to obtain the contribution of a neighborhood point from the map \( B \). For included mask points with few included neighboring points, a low neighborhood contribution value is produced and this results in a high sparsity weight.

In the presence of block inclusion error corruption sparsity alone is not sufficient to give an accurate indication of the likely inclusion errors. Therefore frequency
reliability is also used, where average reliability decision values in each channel are used to predict potential inclusion errors. For a global estimated SNR value $\text{snr}$, the reliability vector $r_{\text{snr}}$ is computed by averaging the reliability decisions over all frames of oracle masks from a validation set:

$$r_{\text{snr}}(f) = \frac{1}{V} \sum_{i=1}^{V} \sum_{t=1}^{K_i} M^i(t, f), \quad \forall f = 1, 2, \ldots, D,$$

where $V$ is the number of validation masks, $K_i$ is the length of the $i$th validation mask, and $r_{\text{snr}}(f) \in [0, 1]$, with values of 1.0 and 0.0 indicating that the observed oracle mask decisions for points in channel $f$ were completely reliable and completely unreliable, respectively. The frequency reliability selection weight for time-frequency point $(t, f)$ is given by

$$w_{\text{FRsnr}}(f) = \frac{1}{r_{\text{snr}}(f) + \epsilon_{\text{FR}}},$$

where constant $\epsilon_{\text{FR}} \in (0, \infty]$ is chosen empirically. The total selection weight for a given TF point is defined as:

$$w(t, f) = w_{\text{sp}}(t, f) \times w_{\text{FRsnr}}(f),$$

$$= M(t, f) \times \frac{1}{r_{\text{snr}}(f) + \epsilon_{\text{FR}}} \times \left[ \sum_{n=n_L}^{n_H} \sum_{q=q_L}^{q_H} M(n, q) \times B(k, l) \right]^{-1}, \quad \text{(21)}$$

Fig. 3. The map $B$ giving the contribution of the neighborhood points within a $5 \times 5$ area to the calculation of the sparsity weight value of the central reference point located at $(t, f)$. 

<table>
<thead>
<tr>
<th></th>
<th>f-2</th>
<th>f-1</th>
<th></th>
<th>f+1</th>
<th>f+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$k$
where the map indices $k$ and $l$ and the neighborhood bounds $n_L$, $n_H$, $q_L$ and $q_H$ are calculated as in (18).

### 4.3. Normalized likelihood distance mask quality

Following group selection, the set of candidate inclusion error points are removed from the mask segment $\mathcal{M}$ to produce mask segment $\mathcal{M}$. Normalized likelihood distances are used to evaluate the quality of the new mask segment $\mathcal{M}$ compared to current segment $\mathcal{M}$. The quality function for mask segment $\mathcal{M}$ of observation segment $\mathcal{X}$ is thus given by

$$O(\mathcal{X}, \mathcal{M}) = \frac{L_{\text{max}}(\mathcal{X} | \mathcal{M}) - E[D_{\text{comp}}(\mathcal{X} | \mathcal{M})]}{L_{\text{max}}(\mathcal{X} | \mathcal{M})},$$

(22)

where $L_{\text{max}}(\mathcal{X} | \mathcal{M})$ and $D_{\text{comp}}(\mathcal{X} | \mathcal{M})$ are the maximal segment likelihood value and the distribution of competitor likelihoods respectively as defined in Sec. 3. Since modification is constrained to the removal of inclusion errors from the initial segment, the global maximum NSLD value over all possible masks within the search space should be produced by the mask segment which removes all unreliable inclusions and thus approximates the oracle segment.

### 4.4. Probabilistic replacement decisions

The refinement of a bottom-up reliability mask segment proceeds via iterative application of the group selection removal and NSLD evaluation steps (see Algorithm 1). Following the calculation of the NSLD for the refined segment $\mathcal{M}$, the replacement procedure is as follows:

- If $O(\mathcal{X}, \mathcal{M}) \geq O(\mathcal{X}, \mathcal{M} \hat{\mathcal{M}})$ then the removal of the candidate unreliable points has increased the mask quality, and thus the current mask is replaced by the refined mask (set $\mathcal{M} = \mathcal{M} \hat{\mathcal{M}}$).
- Otherwise, if $O(\mathcal{X}, \mathcal{M}) < O(\mathcal{X}, \mathcal{M})$ then the mask quality has decreased as a result of the removal, and the refined mask is accepted with a probability $p(t)$ given

$$p(t) = e^{-(t-1)/J}, \quad \text{for } t = 1, 2, \ldots, I,$$

(23)

where $t$ is the current iteration of $I$ total iterations, and $J > 0$ is the time scale controlling the speed with which the acceptance probability approaches 0 with increasing $t$.

Initially, the algorithm accepts the removal of truly unreliable points which decrease the NSLD from the local optimum corresponding to the corrupted input segment, and ensures that mask modifications in latter iterations increase the NSLD towards its global optimum value.
5. Evaluation

5.1. Experimental setup

Experiments were performed on the TIMIT database\textsuperscript{15} to examine the performance of the group selection speaker identification system. The ideal nature of the TIMIT speech makes the database an appropriate choice for robustness experimentation since the exact amount of desired distortion can be added artificially. For all speakers sentences were pre-processed to remove silence, and then segmented into 3 s utterances with segments from the SI and SX sentences used for model training and segments from the SA sentences used for testing. Log-spectral feature vectors were used based on a 48-channel mel-filterbank which operates on 25 ms Hamming windowed frames with a 10 ms frame step. HTR\textsuperscript{46} was used to construct 4-mixture full covariance GMMs to model each speaker.
The test utterances were corrupted with additive white noise from the NOISEX database. Ideal oracle and estimated spectral subtraction TF reliability masks were used in the evaluation. Oracle masks as defined in (16) were subject to random inclusion error corruption such that each noise dominated spectral component was included in the mask with a probability $P_{\text{inc}}$, and to random deletion error corruption such that each speech dominated spectral component was removed from the mask with a probability $P_{\text{del}}$. Spectral subtraction as defined in (17) was also used to demonstrate performance for practical mask corruption, where the SNR threshold $\theta$ was varied between $-3$ dB, 0 dB and 3 dB and results also presented for nonstationary factory noise.

5.2. Baseline missing data experimentation

The performance of a baseline missing data system was first examined to verify the effect of reliability mask errors on the recognition rate. To achieve this 32 TIMIT speakers were chosen by a random selection of four individuals from each of the eight dialect regions. Identification rates were recorded for inclusion and deletion probabilities of $P_{\text{inc}} \in \{0.0, 0.1, \ldots, 1.0\}$ and $P_{\text{del}} \in \{0.0, 0.1, \ldots, 0.9\}$, respectively (see Fig. 4). Random mask corruption produced a decrease in identification rate as the probability of inclusion or deletion error increased. For a given error probability value inclusion corruption produced the most dramatic decrease in recognition performance, particularly in the 10 dB and 0 dB cases. Identification rates remained largely constant for increasing reliable deletion probability until a threshold value was reached, and a sharp degradation was observed beyond this point. This agrees with past research which demonstrates the robustness of marginalization-based missing data to random deletion errors for speech recognition tasks.

Spectral subtraction reliability masking produced recognition rates significantly lower than those of oracle masking, with the best performance obtained using a threshold of $\theta = 3$ dB (Fig. 4(c), lower curves). The removal of inclusion errors in the spectral subtraction masks was achieved by the component wise multiplication of this mask with the oracle mask:

$$m_{\text{tf}}^{\text{oracle}} \times m_{\text{tf}}^{\text{ss}(\theta)}.$$  \hspace{1cm} (24)

Using these inclusion error free masks increased the recognition rate to within 2% of the oracle value for the 20 dB and 10 dB cases. However in the 0 dB case the difference is larger due to the fewer number of truly reliable TF points which increases the effect of the deletion errors in the mask (Fig. 4(c), upper curves).

5.3. Combined system experimentation

To test the correction of random errors in the initial TF mask oracle masks were subject to random inclusion error corruption with speech dominated component inclusion probability $P_{\text{inc}} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Performance was also examined for the use of spectral subtraction mask estimation with $\theta = 0$ dB.
Reliability vector values $r_{\text{snr}}(f)$ were calculated using oracle masks from a set of 10 validation utterances at SNRs within the same 20 dB to −5 dB range as the test speech. Perfect global SNR estimates were used to produce $w_{\text{FR}_{\text{snr}}}$, with $\epsilon_{\text{FR}} = 0.25$ based on separate validation experiments. The NSLD quality measure was calculated with parameters of $K = 50$, $N = 5$, and group selection refinement was performed with the total number of iterations set to $I = 100$. Extensive evaluations over a range of values were carried out to show that segment sizes of at least $K = 50$, consideration of no more than $N = 5$ competitor speaker models, and $I = 100$ iterations would be sufficient. The recognition performance of the proposed system was compared to that of a standard bottom-up only missing data system (i.e. no top-down mask refinement), and also to a missing data system which uses a $5 \times 5$ ‘plus’-shaped median filter$^{25}$ to refine the initial bottom-up masks.

![Graphs showing speaker identification rates](image-url)

Fig. 4. Speaker identification rates for oracle masking with random unreliable inclusion errors (a), oracle masking with random reliable deletion errors (b) and binary spectral subtraction masking (c). Spectral subtraction results include the performance of standard estimated masks (SS) and estimated masks with inclusion errors removed (Oracle*SS).

D. Pullella & R. Togneri
Tuning experiments were first performed to find suitable group size and time scale parameters for each bottom-up mask type. For group sizes of $G \in \{15, 20, 25, 30\}$ and time scales of $J \in \{6, 9, 12, 15, 18, 21\}$, recognition rates were produced on a set of 10 validation utterances. The parameters producing the highest recognition rate on this validation set were noted for each mask type and noise condition.

It is observed that for a given SNR condition the product of the optimal $G$ and $J$ parameters increases for increasing inclusion error corruption inflicted on the oracle mask (see Table 1). This is due to the relationship between the amount of inclusion errors in the initial mask segment, and the number of selections required to remove sufficient mask points such that the NSLD moves away from its initial local optimum and towards the global optimum of the inclusion error free segment. For an accurately estimated mask the number of errors which must be removed from the initial mask segment in order to move away from the local quality function optimum is small. This is reflected in the choice of algorithm parameters such that the acceptance probability for solutions decreasing the quality function rapidly drops for increasing iterations. For poorly estimated masks a larger number of inclusion errors must be removed from the initial segment to escape the local NSLD optimum, and thus larger $G$ and $J$ parameters are chosen such that the acceptance probability remains high for longer.

Using the $G$ and $J$ parameters obtained during tuning, speaker identification experiments were performed for each noise condition and mask type (see Fig. 5). For inclusion error corrupted oracle masking, the use of the group selection enhancement method produced an improvement of up to 65% over the bottom-up only missing data baseline. In comparison to median filtering refinement, iterative group selection produces higher recognition rates in all SNR conditions over all inclusion corruption probabilities. For a low corruption probability the performance difference between the median filtering and iterative group selection strategies remains within 7% at SNRs above $-5$ dB (see Fig. 5(a)). However, when a more severely corrupted initial mask is used iterative group selection refinement provides up to 57% improvement over median filtering which approximates the performance of the bottom-up only

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>SS (0 dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>225</td>
<td>240</td>
<td>300</td>
<td>375</td>
<td>360</td>
<td>300</td>
</tr>
<tr>
<td>15</td>
<td>225</td>
<td>240</td>
<td>300</td>
<td>450</td>
<td>450</td>
<td>450</td>
</tr>
<tr>
<td>10</td>
<td>120</td>
<td>240</td>
<td>270</td>
<td>360</td>
<td>450</td>
<td>420</td>
</tr>
<tr>
<td>5</td>
<td>150</td>
<td>270</td>
<td>225</td>
<td>450</td>
<td>525</td>
<td>360</td>
</tr>
<tr>
<td>0</td>
<td>150</td>
<td>240</td>
<td>300</td>
<td>375</td>
<td>450</td>
<td>180</td>
</tr>
<tr>
<td>−5</td>
<td>120</td>
<td>225</td>
<td>360</td>
<td>420</td>
<td>450</td>
<td>180</td>
</tr>
</tbody>
</table>
Fig. 5. Identification performance for oracle missing data masks with random cell inclusion corruption of
probability \( \mathcal{P}_{\text{inc}} \) (a)–(d) in white noise, and spectral subtraction (SS) estimated masks (e) and (f), in white
and factory noise, respectively. Results are shown for the bottom-up only system using the ideal \textit{a priori}
masks (Oracle Mask), for combined systems using iterative group selection and median filtering refinement
of the corrupted mask (Iterative GS, Median Filtering) and also for the standard bottom-up only system
using the initial mask (Corrupted Mask/SS Mask).
missing data baseline (see Fig. 5(d)). This illustrates the benefit of a recognizer feedback-based decision process to modify the mask segments compared to median filtering which blindly replaces each point by the median of the neighborhood values.

Using spectral subtraction mask estimation the iterative group selection system produced an improvement of up to 32% over the use of only the bottom-up mask, and up to 21% improvement over median filtering-based refinement in white noise (see Fig. 5(e)). For nonstationary factory distortion, the initial subtraction mask estimate is particularly poor, with block error mask corruption often occurring for whole time-frequency regions containing ‘impulse’ noise. In this case, due to the less effective sparsity measure, the group selection algorithm produces lower maximum improvements of 17.19% and 10.93% over the baseline and median filtering strategies respectively (see Fig. 5(f)).

We note that a 0 dB spectral subtraction threshold has been used here for mask estimation. For broadband white noise, increasing this threshold may remove inclusion errors (but increase the risk of deletion errors) in the initial mask estimate resulting in performance improvements for the conventional bottom-up system but relatively less so for the proposed iterative group selection system. However, for nonstationary factory noise increasing the spectral subtraction threshold has less effect on identification performance, and thus the improvement obtained by the proposed method over the conventional bottom-up system is expected to remain large.

The improvements obtained by both the group selection and the median filtering refinement strategies over the baseline are reduced for spectral subtraction masks, compared to corrupted oracle masks. This is due to the decreased effectiveness of TF neighborhood information for choosing candidate inclusion error points for block inclusion corruption compared to random corruption. In this case, while the iterative group selection method outperforms median filtering in high SNRs, at lower SNRs the increased effect of deletion errors (which the proposed approach cannot correct) reduces the improvement obtained.

From Fig. 5, in high SNRs and random inclusion errors the proposed iterative group selection system produced identification rates exceeding (or on par) with that of the oracle mask. Although the oracle mask is considered ideal, in the sense that the mask has perfect estimates of the reliable and unreliable regions, based on SNR thresholding of known speech and noise conditions, this is not strictly the same as stating these masks are upper bounds on the identification rate. For practical cases this is generally true, but in high SNR conditions (as evidenced by Figs. 5(a) and 5(b) especially) this is not necessarily the case.

5.4. Discussion

The results of the evaluation demonstrate the ability of the top-down iterative group selection method to refine estimated bottom-up reliability masks for speaker identification. For corrupted oracle masks and practical spectral subtraction masks a
significant performance improvement is obtained by applying top-down refinement compared to the use of only the initial bottom-up mask. Furthermore, for both masking types iterative group selection enhancement shows considerable improvement over median filtering-based mask post-processing, particularly in the case of poor initial mask estimates with large amounts of inclusion error corruption (see Figs. 5(c) and 5(d)). This demonstrates the successful use of marginal frame likelihoods (via the NSLD) to remove errors within an initial TF reliability mask estimate.

The major motivation for using group selection-based top-down processing is efficiency. Here robustness is provided using global missing data reliability decisions which avoids the calculation of model specific reliable feature subsets, and group selection deletion provides a fast method for the removal of inclusion errors within a corrupted mask segment. However, the approach does have several drawbacks which should be noted. First, the quality of the refinement produced by the proposed algorithm is somewhat dependent on the accuracy with which erroneous points are identified during group selection. A sparsity-based measure is used to perform the selection in this implementation, and this is effective for refining masks where random inclusion errors are the dominant type of corruption. This sparsity measure is, by contrast, much less effective at removing block inclusion corruption which may occur in practically estimated masks. In addition, this implementation of the iterative group selection method only attempts to correct inclusion errors, and the performance improvement is thus limited for initial masks with a large number of deletions.

Second, the approach is dependent on the choice of algorithm parameters used to perform the mask refinement. The group size and time scale determine the number of candidate inclusion errors which are removed from a corrupted mask segment to decrease the NSLD from a local maximum, allowing refinement to proceed towards a segment with global maximum NSLD. The exact number of points which must be removed for this to occur will depend on the noise condition of evaluation and on the bottom-up mask estimation strategy which assigns the reliability decisions. In this implementation the group size and time scale parameters were tuned based on a validation set, and these values were held constant for the refinement of all utterances in the identification test set. However, due to the potentially varying speech energy contents of segments from different utterances, the adjustment of the group size and time scale parameters for the processing of each individual segment should produce even better performance.

Finally, it should be noted that, given the validity of using NSLD-based methods to enhance TF reliability masks, there exists a trade-off between the efficiency and the accuracy with which this refinement is performed. The group selection approach is designed to rapidly traverse the mask search space by exploiting knowledge about which TF points are likely to be inclusion errors. Compared to this method, other techniques which perform a more vigorous examination of the mask search space will possibly achieve a better refinement of the mask segment at the expense of a much larger computation time due to the necessity of calculating segment likelihood scores for each solution considered.
6. Conclusions and Future Work

In this paper we have proposed the combination of bottom-up mask estimation and top-down enhancement for robust missing data recognition. A novel iterative group selection-based method was introduced which utilizes normalized likelihood distances in order to refine initial bottom-up mask estimates independent of a priori noise knowledge. Sparsity and energy criteria are employed to identify potential inclusion errors in the bottom-up mask, and a probabilistic mask quality-based decision process is applied iteratively to remove these errors and thus produce a mask which resembles the oracle mask.

Experimental evaluation was performed on the combined bottom-up top-down system for the task of closed-set text-independent speaker identification. The results showed that combining bottom-up mask estimation and top-down group selection enhancement produced large improvements over the use of a standard bottom-up only missing data system. The results also demonstrated the ability of the proposed group selection approach to significantly outperform a bottom-up top-down combination using median filtering refinement, and this illustrates the benefit of top-down mask enhancement based on marginal frame likelihoods output from the recognizer.

Improving the group selection algorithm is a promising avenue for future work. The algorithm’s dependence on the tuning of the group size and time scale parameters may be reduced by the implementation of a dynamic adaptation technique to automatically adjust these parameters based on the properties of the mask segment being processed. To produce refined masks which more closely resemble the oracle mask optimization methods such as genetic or evolutionary algorithms may be suitable. While these techniques have less dependency on parameter tuning and make fewer assumptions about errors in the initial mask, achieving computational speeds comparable to the group selection approach may be difficult.

Finally, although the top-down approach for missing data mask refinement is presented here as a solution for robust speaker recognition, the generality of the approach extends to any application involving estimation of reliable time-frequency or other image-based masks for recognition. Conceptually, it is also possible to extend the method to handle robust speech recognition. To achieve this the segmentation of the reliability mask must agree with the word boundaries within the corresponding speech utterance being processed. This is to ensure that speech within each individual observation segment corresponds to only one trained HMM; a condition which is necessary for the NSLD criterion and implicitly satisfied in speaker recognition tasks. For the accurate identification of the word boundaries in noisy speech methods proposed in past work may be used.

Acknowledgments

This work was supported in part by the Samaha Research Scholarship (F8046) of The University of Western Australia and the Australian Research Council (ARC) Discovery Project funding scheme (Project Number DP1096348).
References


40. S. Srinivasan and D. Wang, Robust speech recognition by integrating speech separation and hypothesis testing, Speech Commun. 52(1) (2010) 72–81.


43. A. Varga, H. Steenneken, M. Tomlinson and D. Jones, The NOISEX-92 study on the
effect of additive noise on automatic speech recognition, Documentation included in the

44. A. Vizinho, P. Green, M. Cooke and L. Josifovski, Missing data theory, spectral sub-
traction and signal-to-noise estimation for robust ASR: An integrated study, in Proc.

45. R. Weiss and D. Ellis, Estimating single-channel source separation masks: Relevance
vector machine classifiers vs pitch-based masking, Workshop on Statistical and Percep-

46. S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, Hidden
Markov Model Toolkit (HTK) Version 3.2.1 User’s Guide, Cambridge University Engi-
eering Department, Cambridge, MA (2002).

Daniel Pullella completed the B.E. and B. Comp.Sci degrees in 2006
and a Ph.D. in 2011, from
the University of Western
Australia.

From 2005 to 2006, he
was a Research Intern at
the Signal and Informa-
tion Processing Lab, Uni-
versity of Western
Australia. In 2007, he joined the Signals and
Systems Engineering Research Group (SSERG),
University of Western Australia, where he com-
pleted his Ph.D. on robust speaker identification
using time-frequency masking. His research
interests are in the areas of pattern recognition
and signal processing, with specific applications
to robust speaker recognition, noise compensa-
tion for speech processing, and robust speech
feature extraction.

Roberto Togneri (M’89-SM’04) received
the B.E. degree in 1985
and the Ph.D in 1989 both
from the University of
Western Australia. He
joined the School of Elec-
trical, Electronic and
Computer Engineering at
The University of Western
Australia in 1988, where
he is now currently an Associate Professor.

Dr. Togneri is a member of the Signals and
Systems Engineering Research Group and heads
the Signal and Information Processing Lab. His
research activities include signal processing and
robust feature extraction of speech signals, sta-
tistical and neural network models for speech and
speaker recognition, and related aspects of com-
munications, information retrieval, and pattern
recognition. He has published over 80 refereed
journal and conference papers in the areas of
spoken language and information systems, was
co-author of the book Fundamentals of Informa-
tion Theory and Coding Design, Chapman &
Hall/CRC, 2003 and is the chief investigator on
two Australian Research Council Discovery
Project research grants from 2010 to 2013.