Robust Automatic Speech Recognition with Decoder Oriented Ideal Binary Mask Estimation

Lae-Hoon Kim, Kyung-Tae Kim, and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{lkim9,kktae,jhasegaw}@illinois.edu

Abstract

In this paper, we propose a joint optimal method for automatic speech recognition (ASR) and ideal binary mask (IBM) estimation in transformed into the cepstral domain through a newly derived generalized expectation maximization algorithm. First, cepstral domain missing feature marginalization is established using a linear transformation, after tying the mean and variance of non-existing cepstral coefficients. Second, IBM estimation is formulated using a generalized expectation maximization algorithm directly to optimize the ASR performance. Experimental results show that even in highly non-stationary mismatch condition (dance music as background noise), the proposed method achieves much higher absolute ASR accuracy improvement ranging from 14.69% at 0 dB SNR to 40.10% at 15 dB SNR compared with the conventional noise suppression method.

Index Terms: robust speech recognition, ideal binary mask classification, missing feature

1. Introduction

Speech intelligibility is easily degraded by unexpected mismatch. Therefore, speech processing to mitigate or even eliminate the adverse effect of the mismatch has been a very important research topic not only for human-to-human interaction, but also for robust automatic speech recognition (ASR). Note that the source of the mismatch is necessarily neither stationary nor expected. Humans are good at decoding speech even in adverse acoustic environments, e.g., a cocktail party [1, 13]. However, the dependence of ASR on the degree of mismatch is much more critical, and is a very important unsolved problem.

Recently, missing feature theory has been applied to mitigate the negative influence of the mismatch. Missing feature theory classifies the reliable and unreliable feature components first, and uses the reliability information to enhance ASR accuracy [3, 11, 10, 13, 12]. The optimal estimation is rephrased as "imputation" and model adaptation is supplemented with marginalization over missing feature components. Missing feature approach based on complicated spectral domain modeling is reported to be inferior to the cepstral domain modeling with optimal imputation [10]. However, even the optimally enhanced feature is not the same with the original clean feature, and this mismatch is transfered to all the cepstral components, such that it deteriorates the recognition accuracy. Therefore, one of the important current issues is how we can import the missing feature approach into the cepstral domain feature without having to impute a full feature vector from the pre-processing.

Many papers, since the early 1990s, have asked how we can optimally classify the missingness directly for speech recognition performance. Recently, suppressing the missing part via ideal binary mask (IBM) has been demonstrated to be effective for increasing intelligibility of human-to-human speech communication [7]. Typical method for automatic IBM estimation use relatively less complicated front-end processing [3, 11, 7] and the recognition is performed based on the results of the separated pre-processing, which is not necessarily optimal for speech recognition based on a hidden Markov model (HMM).

In this paper, we first introduce how we can use the HMMs trained on the cepstral domain features with a missing feature approach. Our method gives the merits of both: optimal imputation for cepstral domain HMM [10] and missing feature marginalization in spectral domain HMM [3]. Secondly, we introduce decoder-oriented IBM estimation such that we can maximally utilize the information encoded in the parameters of a complete ASR framework. Since the missingness is declared by the speech models, any mismatch different enough from the speech will be classified as unreliable, for example, non-stationary abrupt impulsive noise can be detected and removed. Therefore, limitless cases of mismatch conditions can be handled simply if they are different from the speech used for training the HMMs.

2. Proposed method

2.1. Linear transformation with Gaussian mixture model (GMM)

Speech is contaminated differently in each frequency band. Missing features (features corrupted by mismatch) can be marginalized out, if the HMM has been trained using spectral features [13]. However, an HMM trained on cepstral domain features produces lower word error rate [10]. In this section, this gap is bridged by using the simple fact that the discrete cosine transform (DCT) is a linear transformation of the spectrum, and high order cepstral coefficients are not useful for discriminating among different models. To get the cepstral features, we apply a DCT operation to the spectral coefficients. For example, mel-frequency cepstral coefficients (MFCCs) $\underline{y}$ can be obtained by the following equation.

$$\underline{y} = D\underline{x}, \tag{1}$$

where $\underline{x}$ is mel-frequency spectral coefficients (MFSCs) and D is a DCT matrix such that

$$D(i,j) = \sqrt{\frac{2}{N}} \cos\left(\frac{\pi(i-1)}{N}(j-0.5)\right), \tag{2}$$

where $i, j = 1, 2, \cdots, F$ and F is the number of feature components in a frame. The first row is divided by $\sqrt{2}$ to make the

DCT matrix orthonormal. If we train an HMM with a Gaussian mixture model (GMM)

$$\underline{y}|q, C \sim \sum_{i=1}^{M} \omega_{i|q,C} N \left(\underline{\mu}_{i|q,C}, \text{diag}(\underline{\sigma}^2_{i|q,C}) \right), \quad (3)$$

where C is the target utterance model, q means a specific state and usually we use a diagonal covariance matrix, $\text{diag}(\underline{\sigma}^2_{i|q,C})$. Then,

$$\begin{aligned} \underline{x}|q, C &\sim \sum_{i=1}^{M} \omega_{i|q,C} \\ &\cdot N \left(D^T \begin{bmatrix} \underline{\mu}_{i|q,C} \\ \underline{\mu}_g \end{bmatrix}, D^T \text{diag} \begin{bmatrix} \underline{\sigma}^2_{i|q,C} \\ \underline{\sigma}^2_g \end{bmatrix} D \right) \\ &\sim \sum_{i=1}^{M} \omega_{i|q,C} N(\underline{\mu}'_{i|q,C}, \Sigma'_{i|q,C}), \end{aligned} \quad (4)$$

i.e., the spectral feature distribution is given by a trained HMM with full covariance GMM. Note that it is possible to fill the missing mean and variance for high order coefficients by using global mean vector $\underline{\mu}_g$ and variance vector $\underline{\sigma}^2_g$, independent of the utterance models, states, and mixture components. For example, our experiments use 13 MFCCs transformed from 26 MFSCs. This dimensionality reduction is justified because the remaining coefficients do not play an important role to differentiate the trained models. The replacement with a global mean and variance may be interpreted as a mean and variance tying process across all the different models [2]. Only the mean and variance for the first half of the coefficient vector, multiplied by the number of mixture components, need to be trained using model dependent utterances. Compared with the case when we have to train the full-covariance GMM using spectral features, this is really a huge parameter reduction; at the same time, we are now able to use the missing feature marginalization scheme, since we have an equivalent spectral domain HMMs. Furthermore, we may increase the spectral domain feature dimension, e.g. from 26 to 64 as in [3] if necessary, with relatively small increase of the corresponding cepstral domain. Marginalization of a GMM is performed by simply deleting unreliable components from the mean vector, and deleting the row and column of the unreliable components for the covariance matrix. Therefore, the missing parts will be recognized as "missing" and will not contribute any false information to the acoustic score produced by each model.

2.2. Decoding with missingness classification based on the decoder

In this section, we formulate a joint optimal estimation problem of missingness label (IBM) per each feature component in every frame and ASR, using a generalized EM procedure [8]. Figure 1 shows a conceptual block diagram for the proposed method. Features are labeled as "missing" if doing so increases ASR performance; at the same time, ASR is performed aiming to increase the accuracy of missingness label. For the initialization procedure we may use conventional IBM estimation methods, but we try to rely only on the speech model we already have except for the part of estimating stationary noise level, if any.

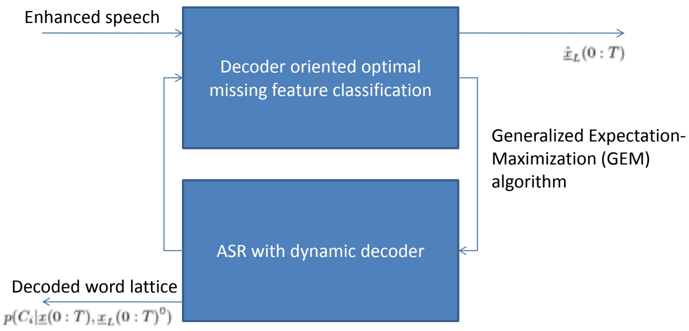
Enhanced speech
Decoder oriented optimal missing feature classification
$\hat{\underline{x}}_L(0:T)$
Generalized Expectation-Maximization (GEM) algorithm
ASR with dynamic decoder
Decoded word lattice
$p(C_i|\underline{x}(0:T), \underline{x}_L(0:T)^0)$

Figure 1: Block diagram for proposed GEM procedure

2.2.1. *Maximum Likelihood Estimation (MLE) of missing feature labels*

IBM estimation produces an estimated binary mask vector sequence (EBMVS),

$$\hat{\underline{x}}_L(1:T) = \arg\max_{\underline{x}_L(1:T)} \log p\left(\underline{x}(1:T)|\underline{x}_L(1:T)\right), \quad (5)$$

where $\underline{x}_L(t, f) \in \{0, 1\}$ according to 1 is for reliable feature, 0 is for unreliable feature, the frame index $t = 1, 2, \cdots, T$ where T is total number of frames, the feature index $f = 1, 2, \cdots, F$ where F is total number of feature. Estimation of $\hat{\underline{x}}_L(1:T)$ can be reformulated into a GEM problem, considering the EBMVS as a parameter to estimate, and the model for ASR as a latent variable.

$$\hat{\underline{x}}_L(1:T) = \arg\max_{\underline{x}_L(1:T)} Q(\underline{x}_L(1:T), \underline{x}_L(1:T)^0), \quad (6)$$

where

$$\begin{aligned} & Q(\underline{x}_L(1:T), \underline{x}_L(1:T)^0) \quad (7) \\ = \ & E_{p(C|\underline{x}(1:T), \underline{x}_L(1:T)^0)} \left[\log p(\underline{x}(1:T), C|\underline{x}_L(1:T)) \right]. \end{aligned}$$

In the maximization step, we need to solve the following equation to estimate $\hat{\underline{x}}_L(t, f)$.

$$\begin{aligned} & \hat{x}_L(t, f) \\ = \ & \arg\max_{x_L(t,f) \in \{0,1\}} \sum_{i=1}^{N_C} p(C_i|\underline{x}(1:T), \underline{x}_L(1:T)^0) \\ \cdot \ & \sum_q \left[\sum_{q'} \alpha_{t-1}(q'|C_i, \underline{x}(1:t-1), \underline{x}_L(1:t-1)^0) p(q|q') \right] \\ \cdot \ & \beta_t(q|C_i, \underline{x}(t+1:T), \underline{x}_L(t+1:T)^0) \quad (8) \\ \cdot \ & \log p(\underline{x}(t)|x_L(t, f), \underline{x}_L(t, \neg f) = \underline{x}_L(t, \neg f)^0, C_i, q), \end{aligned}$$

where C_i represents an utterance model and N_C is the total number of utterance models, $\underline{x}_L(t, \neg f)$ represents $\underline{x}_L(t)$ excluding $\underline{x}_L(t, f)$, $\underline{x}_L(1:T)^0$ is the estimate of EBMVS at previous iteration, and $\alpha_t(q)$ and $\beta_t(q)$ are the conventional forward and backward variables at time t and state q [9].
Instead of following (8), in this paper we replace the marginalization for state sequence by forward-backward algorithm with the best sequence by Viterbi decoding as shown in (9). Then, the expectation step will be less computationally expensive and the maximization step can also be implemented with less for-loops, because we are given the states at each frame, so all frames can

be independently updated.

$$\begin{aligned}
&\hat{x}_L(t,f)\\
=\;& \underset{x_L(t,f)\in\{0,1\}}{\arg\max} \sum_{i=1}^{N_C} p(C_i|\underline{x}(1:T),\underline{x}_L(1:T)^0) \qquad (9)\\
\cdot\;& \log p(\underline{x}(t)|x_L(t,f),\underline{x}_L(t,\neg f)=\underline{x}_L(t,\neg f)^0,C_i,q_t^{C_i}),
\end{aligned}$$

where $q_t^{C_i}$ represents a best path state at time t in a given model C_i. For fair comparison, some normalization procedure is needed to prevent the situation where we always choose $x_L(t,f)=0$ in (9). To prevent this, we may want to claim that we should know a mismatch source (noise) model responsible for the missing components. However, because often times it is not available, we introduce a method to resolve this issue without necessity of those noise models in the following.

$$\hat{x}_L(t,f) = \underset{x_L(t,f)\in\{0,1\}}{\arg\max} f(x_L(t,f)), \tag{10}$$

where $f(1)$ is the same as (9), but $f(0)$ is formulated as follows:

$$\begin{aligned}
&f(0)\\
=\;& \sum_{i=1}^{N_C} p(C_i|\underline{x}(1:T),\underline{x}_L(1:T)^0) \qquad (11)\\
\cdot\;& [\log p(\underline{x}(t,\neg f)|\underline{x}_L(t,\neg f)=\underline{x}_L(t,\neg f)^0,C_i,q_t^{C_i})\\
+\;& \log p_n(x(t,f)|\underline{x}_L(t,\neg f)=\underline{x}_L(t,\neg f)^0,C_i,q_t^{C_i})],
\end{aligned}$$

where

$$\begin{aligned}
&p_n(x(t,f)|\underline{x}_L(t,\neg f)=\underline{x}_L(t,\neg f)^0,C_i,q_t^{C_i}) \qquad (12)\\
=\;& \underset{x(t,f)\in\mathrm{CI}_\alpha}{\arg\max}\; p(x(t,f)|\underline{x}_L(t,\neg f)=\underline{x}_L(t,\neg f)^0,C_i,q_t^{C_i}).
\end{aligned}$$

In statistics, a confidence interval (CI) is an interval estimate of a random parameter [14]. Probability that the interval includes the parameter is determined by the confidence level α. For example, if we simplify (12) to the case of 2 reliable feature components $x(t,f_1)=a$, $x(t,f_2)=b$ in a given model C and state q, then $p_n(x(t,f)|x(t,f_1)=a,x(t,f_2)=b,C,q)$ can be described as follows.

$$p_n(x(t,f)|x(t,f_1)=a,x(t,f_2)=b,C,q) \sim N(\mu,\Sigma), \tag{13}$$

where

$$\begin{aligned}
\mu \;=\;& E[x(t,f)] + \mathrm{Cov}[x(t,f),[x(t,f_1)\;x(t,f_2)]] \qquad (14)\\
\cdot\;& \mathrm{Cov}[[x(t,f_1)\;x(t,f_2)]]^{-1}([a\;b]-E[x(t,f_1)\;x(t,f_2)])^T,
\end{aligned}$$

and

$$\begin{aligned}
\Sigma \;=\;& \mathrm{Cov}[x(t,f)] - \mathrm{Cov}[x(t,f),[x(t,f_1)\;x(t,f_2)]]\\
\cdot\;& \mathrm{Cov}[[x(t,f_1)\;x(t,f_2)]]^{-1}\\
\cdot\;& \mathrm{Cov}[[x(t,f_1)\;x(t,f_2)],x(t,f)]. \qquad (15)
\end{aligned}$$

2.2.2. Initialization

Before starting iterative IBM estimation, a reasonably good starting point with GEM is important. The following schemes are designed to achieve a good starting point.

<**Stationary background noise filtering**>
We can easily obtain the stationary background noise statistics [4], therefore spectral components below the stationary noise floor can be initially labeled as “missing”’ $(x_L(t,f)^0=0)$.

	MFCC	MFSC	MFSC-missing one
Accuracy(%)	94.24	94.24	94.64

Table 1: Recognition accuracy(%)

<**Confidence Interval**>
The second part of the initialization scheme is formalized as follows:

$$H_0 : x \in \cup_{\theta\in\Theta}\mathrm{CI}_{\alpha_{high},\alpha_{low}|\theta} \tag{16}$$

$$\mathrm{CI}_{\alpha_{high},\alpha_{low}|\theta} = (x_L, x_H), \tag{17}$$

such that $P(x < x_L) = \alpha_{low}$ and $P(x > x_H) = \alpha_{high}$, where Θ includes all states in all models. Basic object of this scheme is to detect the missing parts based on confidence scores of speech models, which have been already trained in a target environment, e.g. high SNR environment.

<**Speech model based voice activity detection (VAD)**>
At the end of the initialization process, we apply a model based VAD, which is dependent on the relative number of reliable components in each frame compared with the maximum number of reliable components among all frames in a given utterance. We declare a frame with fewer reliable components than the threshold as missing, thereby removing frames with very few reliable feature components.

3. Experiment

To validate the proposed method, isolated digit HMM recognizers with a single Gaussian per state were built using 13 MFCCs transformed from 26 MFSCs. HTK [2] was used to train the 12 different models “one”, “two”, $\cdots$, “nine”, “zero”, “oh” and “silence” using the TIDigits corpus [6]. After training the models, we transformed the models back to the spectral domain by following (4). When we train HMM models, we can exclude the silence model using forced alignment. Given only the digit models, the silence frames (presumably contaminated with background noise) can be classified as missing.

Three different tests were performed to check the validity of (4), first using cepstral features, second using spectral features for the testing data, and third using part of the spectral features to simulate the missing features case. Note that for the first set, we just need to use the original HMM model, trained on 13 MFCCs. For the second and third sets, we need to use the linearly transformed HMM models. As discussed previously, the global mean and variance have been used to fill the last 13 high order MFCCs. For the third case, the 26th MFSC, which summarizes the energy between 3.6 kHz and 4.0 kHz, is simulated as missing. Table 1 shows the recognition accuracy in all three cases. As expected, the transformed HMM produces exactly the same accuracy as the original HMM. Interestingly, the accuracy of clean-speech TIDigits recognition is improved slightly when we marginalize out the last MFSC.

Figure 2 shows the ASR accuracy of the noisy speech with white noise as background noise, after applying enhancement by MMSE logSA [4], and after MMSE logSA together with the proposed scheme. Note that MMSE logSA with the proposed method outperforms MMSE logSA alone. Figure 3 (highly non-stationary dance music as background noise) shows more interesting results with the proposed scheme. A conventional

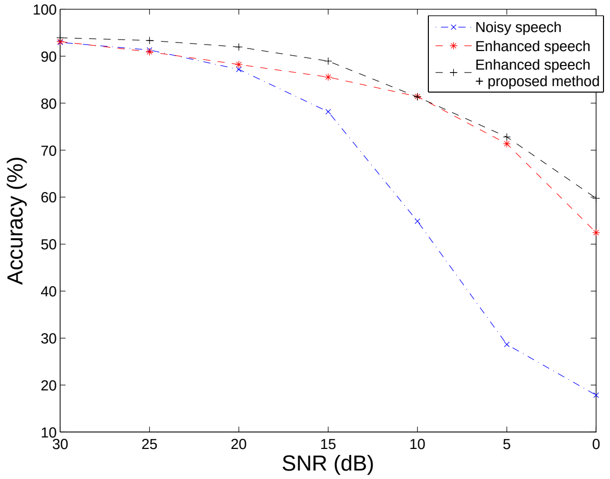

Figure 2: Recognition accuracy(%) vs SNR (dB) with white noise as background noise

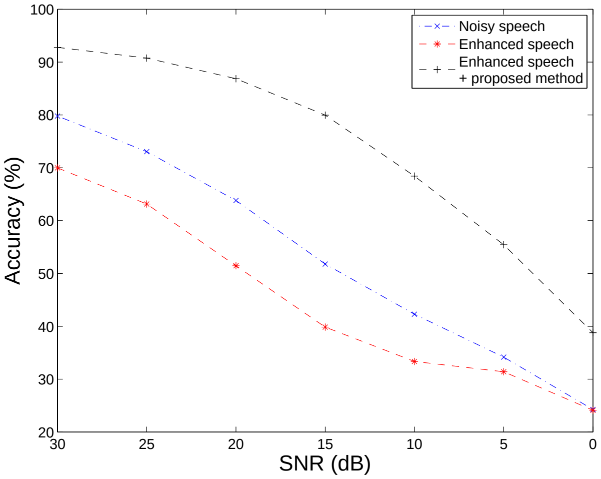

Figure 3: Recognition accuracy(%) vs SNR (dB) with dance music in the movie "Dream Girls" as background noise

speech enhancement algorithm deteriorates the ASR performance. However, the proposed method highly outperforms the baseline. Normally it is known to be very hard to achieve robust speech recognition in music-like non-stationary background noise [11], but the proposed method is performing much better than baseline in this kind of unexpected, untrainable mismatch condition.

4. Conclusion and future work

This paper uses the knowledge of a trained ASR to classify the reliable feature components for missing-feature speech recognition, which in turn contributes to increase the accuracy of ASR. Signal enhancement method like MMSE logSA can be combined with the proposed scheme without hurting the performance of the front-end processing. In fact, if we can have more reliable feature component through this separate front-end processing, it supplies us a better chance to have more accurate speech recognition results. GEM iteration iteratively approaches the utterance specific speech process models allowing better estimation of the binary mask. To summarize, the proposed method can provide better ASR performance not only for the case of stationary mismatch (by combining with preprocessing), but also for the case of the non-stationary mismatch, where the mismatch is not easily modeled or estimated. In the experimental study on understanding the cocktail part effect, the binaural aspect of human hearing has been emphasized. This location based multichannel information will be combined to boost the IBM estimation accuracy with the proposed one-channel scheme given the correct estimate of the direction of arrival (DOA) of the target speech per each feature index [5].

5. References

[1] B. Arons. A review of the cocktail party effects. In *MIT Media Lab. Retrieved on December 18, 2006*, 1992.

[2] Cambridge University Engineering Department. *HTK Speech Recognition Toolkit*, January 2007.

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285, 2001.

[4] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33:443–445, April 1985.

[5] L.-H. Kim, M. Hasegawa-Johnson, G. Potamianos, and V. Libal. oint estimation of doa and speech based on em beamforming. In *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2010.

[6] R. G. Leonard. A database for speaker-independent digit recognition. In *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 1984.

[7] Y. Li and D. Wang. On the optimality of ideal binary time-frequency masks. *Speech Communication*, 51(3):230 – 239, 2009.

[8] R. M. Neal and G. E. Hinton. *A view of the EM algorithm that justifies incremental, sparse, and other variants in M. I. Jordan. ed Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[9] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceeding of the IEEE*, 77:257–285, 1989.

[10] B. Raj, M. L. Seltzer, and R. M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43:275–296, 2004.

[11] M. L. Seltzer, B. Raj, and R. M. Stern. A bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 43:379–393, 2004.

[12] S. Srinivasan and D. Wang. Transforming bianary uncertainties for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:2130–2140, 2007.

[13] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis Principles, Algorithms, and Applications*. Springer, Berlin, 2006.

[14] S. Weisberg. *Applied Linear Regression, Third edition*. Wiley, New Jersey, 2005.