7th International Conference on Advanced Computational Intelligence
Mount Wuyi, Fujian, China; March 27-29, 2015

Noise Adaptive Stream Fusion Based on Feature Component Rejection for Robust Multi-Stream Speech Recognition

Jun Zhang, Yizhi Feng, Gengxin Ning, and Fei Ji

***Abstract*—Weighting the stream outputs according to their reliability levels is one of the most common stream fusion methods in the multi-stream automatic speech recognition (MS ASR). However, when a MS ASR system works in noisy environments, there are distortion level differences among not only the data streams, but also the feature components inside a stream. In this paper, we first propose a feature component rejection approach that can provide the similar function as the missing data techniques while is much easier to be applied to different features. Then a new stream fusion method that can make use of the reliability information of both inter- and intra-streams is developed by incorporating the proposed feature component rejection approach into the conventional MS HMM. The proposed stream fusion method shows good noise adaptive ability and achieves similar recognition accuracy as the missing data based stream fusion method for additive noises in the experiments of the Ti digits connected word recognition task.**

I. Introduction

The multi-stream (MS) approaches have been wildly used in the automatic speech recognition (ASR) for many years and proven to be effective ways to improve the recognition accuracy and robustness of the ASR systems. Based on the fact that the noises or mismatches do not affect the different data streams in similar ways, the MS recognizers can usually outperform the single stream ones in various and unpredictable noisy environments by choosing and fusing the complementary data streams properly. Many combinations of feature streams were proposed in the previous researches and gave encourage results, such as the combination of the audio and visual streams [1], the standard and throat microphone streams [2], and the complementary acoustic streams [3].

Usually different data streams are merged by weighting the stream outputs according to their reliabilities at a certain level of the recognition. However, while the noises cause unequal distortions among data streams, they also affect the feature components inside a stream differently. The missing data theory shows that if the reliability differences of the feature components are taken into account, the robustness of the recognizer can be improved significantly [4]. Obviously, the conventional MS framework that assigns weights merely on the stream outputs cannot make use of this “intra-stream” reliability information. In our previous work [5], missing data techniques were proposed to incorporate into the conventional MS fusion framework to handle these intra-stream reliability differences and provide the recognizers noise adaptive abilities. However, these methods have the disadvantage that the explicit reliable/unreliable data detection modules required by the missing data techniques are feature type dependent and difficult to implement for many kinds of features that will be used in the MS ASR systems.

In this paper, we propose a new stream fusion method to make use of the reliability information of both inter- and intra-streams. First, we extend the idea of out-of-vocabulary (OOV) word rejection to the feature component level to develop a feature component rejection approach which can provide the similar function as the missing data techniques while is much easier to be applied to different features. Then this new approach is incorporated into the conventional MS HMM to deal with the intra-stream distortion differences and provides an efficient way to adjust the output sensitivity of the MS recognizer to each data stream adaptively. Experiments on the TI digits connected word recognition task are carried out to evaluate the performance of the proposed stream fusion method.

II. Feature Component Rejection

The OOV word rejection [6]-[8] has been used in many practical ASR systems to improve the robustness by detecting and ignoring the unknown words, whose basic idea is quite similar to the missing data technique but works at the word level and is usually feature type independent. Therefore it is attractive to extend the idea of OOV word rejection to the feature component level to handle the unknown feature components caused by distortions.

In many small vocabulary speech recognition tasks such as voice control and keyword spotting, the OOV word rejections are often implemented by representing the OOV words with a set of garbage models and rejecting the words whose confidences are lower than the thresholds that are determined by the garbage scores [7], [8]. To extend this technique to the feature component level, consider the multi-Gaussian HMM (MG HMM) with diagonal covariance which is wildly used in the ASR. Let $\mathbf{o}=\left(o_1,o_2,\cdots,o_I\right)$ denotes the observed feature vector with the dimension of I . In this letter, we model the 'OOV' distribution of o_i by a single explicit garbage model $p^g\left(o_i\right)$ which is independent of the HMMs and their states. Because the feature components in each Gaussian mixture of the diagonal MG HMM can be regarded as independent, we

Manuscript received December, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant numbers 61372083 and 61372080, and the Guangzhou Science and Technology Project 2012J2200005.

The authors are with the School of Electronic and Information Engineering, South China University of Technology, Guangdong, China. (e-mail: eejzhang@scut.edu.cn; yzfeng@scut.edu.cn; ninggx@scut.edu.cn; eefeiji@scut.edu.cn).

modify the output probability of the MG HMM by applying the garbage model to each mixture as

$$p^{mg}\left(\mathbf{o},\mathbf{v}\,|\,\lambda\right)=\sum_{m=1}^{M}a_m\prod_{i=1}^{I}\left(v_i p_m^c\left(o_i\,|\,\lambda\right)+\left(1-v_i\right)p^g\left(o_i\right)\right), \quad (1)$$

where λ is a given state in the HMM, M is the number of mixture, a_m is the weight of the mth mixture, $p^{mg}\left(\mathbf{o},\mathbf{v}\,|\,\lambda\right)$ and $p_m^c\left(o_i\,|\,\lambda\right)$ denote the output probability of the MG HMM and the clean Gaussian distribution (i.e., the in-vocabulary (IV) distribution) of o_i in the mth mixture respectively, $v_i\in\left[0,1\right]$ denotes the rejection degree of o_i and $\mathbf{v}=\left(v_1,v_2,\cdots,v_I\right)$. By introducing the garbage model and v_i into the MG HMM, (1) provides an efficient way to adjust the output probability of each feature component. There are two major problems needed to be considered when using (1), i.e., the calculation of $\mathbf{v}$ and the design of garbage models.

A. Calculation of $\mathbf{v}$

The requirement of feature type dependent reliable/unreliable detection module is the major drawback of the missing data techniques when they are applied to the MS ASR. To overcome this problem, we develop a state based maximum likelihood (ML) estimator for $\mathbf{v}$ by using the garbage model as

$$\mathbf{v}^{ml}=\arg\max_{\mathbf{v}} p^{mg}\left(\mathbf{o},\mathbf{v}\,|\,\lambda\right), \quad (2)$$

where $\mathbf{v}^{ml}=\left(v_1^{ml},v_2^{ml},\cdots,v_I^{ml}\right)$ is the ML estimation of $\mathbf{v}$. Substitute (1) into (2), the ith component in $\mathbf{v}^{ml}$ can be calculated by

$$v_i^{ml}=\begin{cases}1, & if\ \sum_{m=1}^{M}a_m p_m^c\left(o_i\,|\,\lambda\right)>p^g\left(o_i\right)\\ 0, & otherwise\end{cases}. \quad (3)$$

According to (3), the calculation of $\mathbf{v}^{ml}$ is only relied on the IV and OOV speech models, which is irrelevant to the feature type used in the frontend.

B. Garbage Model for the Feature Component

The design of garbage models is another essential problem to the proposed approach. Because the OOV words are usually unpredictable and absent of training data, some garbage models were obtained from the IV data or online IV scores in the previous studies [7-8]. In these cases, they were not required to model the distributions of the OOV words very accurately, but needed to be effective in rejecting the words with low confidences (for example, low IV scores). Following this idea, we use a simple garbage model to reject the components with low IV probabilities by assuming that the OOV components will have uniform distributions associated with the range of values observed during training [9-10] as

$$p^g\left(o_i\right)=\alpha\left(o_i^{\max}-o_i^{\min}\right)^{-1}, \quad (4)$$

where $o_i^{\max}$ and $o_i^{\min}$ are the maximum and minimum values of the ith component observed in the training data. a is a scaling factor for adjusting the rejection thresholds. The components with values outside $\left(o_i^{\min},o_i^{\max}\right)$ are regarded as OOV so that their output probabilities are calculated by (4) as well.

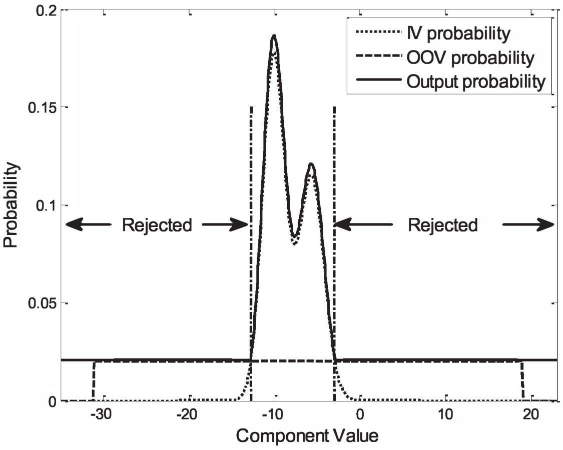

Fig. 1. An example of the rejection region and output probability of a MFCC component. The output probability is elevated 1% for better viewing. A component with the marginal IV probability lower than the OOV probability will be rejected and output the same OOV probability for each candidate state.

According to (3), a component with the marginal IV probability lower than (4) will be rejected and output the garbage probability which is equal for all candidate states, as shown in Fig.1. It is reasonable to assume that a severely distorted feature component will be more likely to have a lower marginal IV probability than the OOV probability for most candidate states if the garbage models are properly designed. So it will tend to be rejected and ignored by the recognizer like in the missing data techniques because in this case its output scores for most candidate states will be equal so that provide no discriminative information, i.e., the proposed approach can be expected to provide similar performances as the missing data techniques.

C. Comparing with Missing Data Technique

Although (1) has a similar form to the missing data technique proposed in [10], there is significant difference between these two approaches on the calculation of $\mathbf{v}$. In [10], $\mathbf{v}$ is interpreted as the missing mask vector which is calculated by a state independence estimator based on the feature domain SNR before the Viterbi decoding. The difficulty in evaluating the feature domain distortion brings great limitation to the choices of features. In the proposed approach, $\mathbf{v}$ is determined by the state based ML estimator during the Viterbi decoding without explicit estimation on the distortions of the feature components or any other feature type specified information. Therefore it is much easier to be applied to different features and more suitable for the MS scenarios.

III. Stream Fusion based on Feature Component Rejection

Assume that the observation $\hat{\mathbf{o}}=\left(\mathbf{o}_1,\cdots,\mathbf{o}_J\right)$ includes J

data streams, where $\mathbf{o}_j$ is the observation of the *j*th stream. The conventional stream fusion methods regard the estimation of the output score $S^{ms}\left(\hat{\mathbf{o}},\mathbf{w}\mid\lambda\right)$ in the MS recognizer as a combination of the single stream score $S_j^{ss}\left(\mathbf{o}_j\mid\lambda\right)$ with a fusion function [11] as

$$S^{ms}\left(\hat{\mathbf{o}},\mathbf{w}\mid\lambda\right)=f\left(\mathbf{w},S_1^{ss}\left(\mathbf{o}_1\mid\lambda\right),\cdots,S_J^{ss}\left(\mathbf{o}_J\mid\lambda\right)\right), \qquad (5)$$

where $\mathbf{w}=\left(w_1,\cdots,w_J\right)$ is the weighting vector for the streams, $f(\cdot)$ is the fusion function. The output score can be probability, logarithm probability, and etc.

In (5), $\mathbf{w}$ can be used to control the outputs of the data streams according to their reliability levels. However, when a MS recognizer works in a noisy environment, unequal distortions exist not only among data streams, but also the feature components inside a stream. The previous researches on missing data theory have showed that if the reliability differences of the feature components are taken into account, the robustness of the recognizer can be improved significantly. In the conventional MS framework represented by (5), all feature components inside a data stream share the same stream weight so that the MS recognizer has no ability to make use of this intra-stream reliability information. In this paper, we develop a new stream fusion model to handle this problem.

A. Framework

To make use of the intra-stream reliability information, a new MS fusion model can be developed by introducing a vector $\hat{\mathbf{v}}=\left(\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_J\right)$ that is associated with the reliabilities of all feature components into the fusion function as

$$S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)=f\left(\mathbf{w},S_1^{ss}\left(\mathbf{o}_1,\mathbf{v}_1\mid\lambda\right),\cdots,S_J^{ss}\left(\mathbf{o}_J,\mathbf{v}_J\mid\lambda\right)\right) (6)$$

where $\mathbf{v}_j$ is associated with the reliabilities of the feature components in the *j*th stream. Comparing with the conventional stream fusion methods, $\hat{\mathbf{v}}$ can be used to adjust the outputs of the feature components according to their reliability levels, which makes the fusion model of (6) has the potential ability to outperform the conventional ones.

In (6), both $\mathbf{w}$ and $\hat{\mathbf{v}}$ can affect the stream output, so a new measurement should be developed to evaluate the influence of $\mathbf{o}_j$ on $S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)$. Generally, the influence of a data stream on the decision is determined by the sensitivity of the output score to the change of the feature vector in this stream [5]. Therefore, the power or module of the directional derivative of $S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)$ on $\mathbf{o}_j$, which determines the sensitivity of $S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)$ to the change of $\mathbf{o}_j$, can be used to evaluate the influence of each data stream on the decision. So for a given $\mathbf{w}$ and $\hat{\mathbf{v}}$, we define a new measurement called the Output Sensitivity to a Stream (OSS) to evaluate the influence of $\mathbf{o}_j$ on $S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)$ as

$$\begin{aligned}OSS_j&=E\left[\left|\frac{\partial S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)}{\partial\mathbf{o}_j}\right|^2\Big/\sum_{l=1}^{J}\left|\frac{\partial S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)}{\partial\mathbf{o}_l}\right|^2\right]\\&=E\left[\left|\frac{\partial S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)}{\partial\mathbf{o}_j}\right|^2\Big/\left|\nabla S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)\right|^2\right]\end{aligned}, \qquad (7)$$

where $\nabla S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)$ is the gradient of $S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)$ on $\hat{\mathbf{o}}$. A bigger OSS_j indicates that the *j*th stream will have more influence on the output score with the current setting of $\mathbf{w}$ and $\hat{\mathbf{v}}$, and vice versa. So the OSS vector can be used to analysis the influence of a data stream to the decision in the new stream fusion model theoretically.

B. Stream Fusion with Feature Component Rejection

Combining the feature component rejection approach with the conventional MS HMM which merges the streams in the state level by the widely used exponential combination [12], a new stream fusion model can be given by

$$p^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)=\prod_{j=1}^{J}\left(p_j^{mg}\left(\mathbf{o}_j,\mathbf{v}_j\mid\lambda\right)\right)^{w_j}, \qquad (8)$$

where $p_j^{mg}\left(\mathbf{o}_j,\mathbf{v}_j\mid\lambda\right)$ is the output probability of the *j*th stream which is calculated by (1), $p^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)$ is the output possibility of the MS HMM.

To exam the effect of rejecting a feature component on the decision, assume that the logarithm probability is used as the output score, i.e., $S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)=\log p^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)$. Substituting (1), (3), (4) and (8) into (7), it can be easy to prove that

$$OSS_j\Big|_{v_{ji}=1}>OSS_j\Big|_{v_{ji}=0} \qquad (9)$$

when $w_j\neq 0$ and $\left|\frac{\partial S^{ms}\left(\hat{\mathbf{o}},\mathbf{w},\hat{\mathbf{v}}\mid\lambda\right)}{\partial\mathbf{o}_j}\right|$ are not all zeros, where v_{ji} denotes the rejection degree of the *i*th component in the *j*th stream.

(9) means that when a component in a data stream is rejected, the influence of this stream to the decision will be reduced. Therefore, the more feature component in a data stream is rejected, the less sensitive the output score of the MS recognizer shows to this stream. Generally, the components in a severely distorted data stream will be more likely to be rejected than in a less distorted stream if the garbage models are properly designed. So the recognizer will tend to make its decision more relied on the less distorted streams, i.e., combining the feature component rejection approach into the conventional MS HMM provides an efficient way to adjust the output sensitivity to each data stream adaptively.

IV. Experiments

The proposed method is evaluated on the TI Digits speaker-independent connected-digit corpus and NOISEX92 noise database. 13 word-level HMMs (1-9, 'oh', 'zero', a silence and a short pause model) are trained on the adult part of the training section in the TI Digits corpus, which contains 8623 utterances from 55 males and 57 females. Each digit HMM composes of 13 no-skip straight-through emitting states with 3 diagonal Gaussian mixtures per state. The silence HMM has 3 full connected states with 6 diagonal Gaussian mixtures per state. The short pause HMM is a 1 state tee-model which has its emitting state tied to the center state of the silence model. The testing set includes 770 utterances from 5 males and 5 females randomly selected from the TI Digits testing section. The white noise, F16 noise, and factory noise from the NOISEX92 database are artificially added to the clean speeches with different SNRs. The HTK tools [12] are used in both the model training and recognition.

Two data streams extracted from the acoustic speech signal by different methods are used in the experiment. The first stream consists of 12 full band MFCC coefficients extracted from the log energy outputs of 26 Mel filter banks. The second stream consists of 12 sub-band MFCC coefficients with 6 MFCCs extracted from a lower sub-band of 0Hz ~ 2178Hz (i.e., the 13 lower filter bands) and 6 MFCCs extracted from an upper sub-band of 1902Hz ~ 10000Hz (i.e., the 13 upper filter bands) [5]. The delta coefficients are computed and appended to the basic acoustic vectors. The sub-band feature is known to be robust against band-limited and some color noises, but not as efficient as the full band feature for many full band noises. So it can provide complementary information to the full band feature. The conventional MS HMM recognizer provided by [12] is used as the baseline system.

A. Comparison of the Conventional and the Proposed Methods

To demonstrate the effectiveness of the proposed method, the recognition accuracy of the proposed approach is compared with two MS approaches under different types of noises with the SNR from 5dB~20dB. The first is the baseline system with linear spectral subtraction, which adopts the conventional MS HMM framework that weights only the data streams. The weights of both streams are empirically set to 0.5, which can achieve average good performance for different noises in the experiment. The second is the missing data based stream fusion method [5] with the soft decision missing data approach proposed in [10], in which weights are assigned to both data streams and feature components. In our previous research, the latter showed good noise adaptive ability and significant improvement over the conventional stream fusion method, but has the disadvantage of requiring feature type dependent reliable/unreliable data detection modules. For better comparison, both stream weights are set to 0.5 and the same linear spectral subtraction are adopted in the missing data based stream fusion method and the proposed method as well. The HTK recognition tools are modified according to (1), (3), (4) and (8) to implement the proposed method. α in (4) is empirically set to 0.2.

Table I shows the recognition accuracy of the baseline system, the baseline system with linear spectral subtraction, the missing data based stream fusion method and the proposed method in different noisy environments. It can be seen that both the proposed method and the missing data based stream fusion method outperform the conventional MS HMM significantly. This result shows that using both inter- and intra-stream reliability information can improve the robustness of the MS recognizer effectively. The proposed method achieves similar recognition accuracy as the missing data based stream fusion method in different noises but without feature type dependent weight estimation for the feature components, which makes it more suitable for the MS scenarios. However, the recognition accuracy of the proposed approach will slightly decrease for clean speeches due to some mis-rejection of the clean feature components which have low IV probabilities. The performance will also drop when the SNR is low due to the scarcity of the un-rejected components.

TABLE I
RECOGNITION ACCURACY OF DIFFERENT METHODS (%)

(a) White Noise

SNR (dB)	∞	20	15	10	5
Baseline	99.1	64.6	42.6	21.8	11.9
Baseline with spectral subtraction	99.1	93.6	88.2	72.1	42.6
Stream fusion based on soft decision	97.7	96.6	92.5	81.1	55.8
Stream fusion based on feature component rejection	98.4	96.1	93.6	83.1	60.2

(b)F16 Noise

SNR (dB)	∞	20	15	10	5
Baseline	99.1	96.3	89.7	65.3	29.3
Baseline with spectral subtraction	99.1	92.6	91.0	84.5	64.5
Stream fusion based on soft decision	97.7	97.6	96.1	90.7	72.8
Stream fusion based on feature component rejection	98.4	98.1	96.0	90.0	73.2

(c) Factory Noise

SNR (dB)	∞	20	15	10	5
Baseline	99.1	96.3	87.4	60.6	26.8
Baseline with spectral subtraction	99.1	87.5	86.6	81.7	64.3
Stream fusion based on soft decision	97.7	97.4	95.0	88.5	67.5
Stream fusion based on feature component rejection	98.4	97.2	95.4	90.6	72.4

TABLE II
RECOGNITION ACCURACY OF SINGLE AND MULTI STREAMS (%)

(a) White Noise

SNR (dB)	∞	20	15	10	5
Full band stream	98.6	95.5	91.6	79.4	54.6
Sub-band stream	98.0	95.6	89.5	74.7	48.1
Multi-stream	98.4	96.1	93.6	83.1	60.2

(b)F16 Noise

SNR (dB)	∞	20	15	10	5
Full band stream	98.6	97.1	94.9	88.7	70.8
Sub-band stream	98.0	97.5	96.1	91.6	73.0
Multi-stream	98.4	98.1	96.0	90.0	73.2

(c) Factory Noise

SNR (dB)	∞	20	15	10	5
Full band stream	98.6	97.1	94.3	87.7	70.2
Sub-band stream	98.0	96.6	93.4	86.2	63.5
Multi-stream	98.4	97.2	95.4	90.6	72.4

B. Noise Adaptive Ability

To show the noise adaptive ability of the proposed method, the recognition accuracy of the proposed method is compared with the single full band and sub-band stream recognizers. The full band recognizer and the sub-band recognizer are obtained by setting the correspondent stream weight to 1 and the other stream weight to 0 under the MS framework. In the MS recognizer, both weights of the two data streams are set to 0.5.

Table II shows the recognition accuracy of the full band stream, the sub-band stream and the multi-stream fused by the proposed method. From table II it can be seen that the full band and sub-band feature streams give different performances when the type of environment noise changes. So using only one data stream isn't a good choice for a speech recognizer that has to face various kinds of noises. On the other hand, the proposed MS recognizer show good noise adaptive ability even with a fixed w by giving similar recognition accuracy as or outperforming the best single stream recognizer when the environment noise changes.

V. Conclusion

In this paper we propose a new stream fusion method based on feature component rejection. The proposed method shows good noise adaptive ability and achieves similar recognition accuracy as the missing data based stream fusion method while can be applied to different features more easily. Further improvement can be expected by using more sophisticated garbage models and more accurate decision on the OOV components.

References

[1] V. Estellers, M. Gurban, and J. Thiran, “On Dynamic Stream Weighting for Audio-Visual Speech Recognition,” in *Proc. IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, pp. 1145 – 1157, 2012.

[2] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, “Combining standard and throat microphones for robust speech recognition,” *IEEE Signal Processing Letters*, vol. 10, pp. 72-74, 2003.

[3] S. K. Nemala, K. Patil, and M. Elhilali, “A Multistream Feature Framework Based on Bandpass Modulation Filtering for Robust Speech Recognition,” in *Proc. IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, pp. 416-426, 2013.

[4] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, “Robust automatic speech recognition with missing and unreliable acoustic data,” *Speech Communication*, vol. 34, pp. 267-285, 2001.

[5] J. Zhang, G. Wei, H. Yu, and G. X. Ning, “Robust multi-stream speech recognition based on weighting the output probabilities of feature components,” *Chinese Journal of Acoustics*, vol. 28, pp. 269-279， 2009.

[6] B. Rveil, K. Demuynck and J. P. Martens, “An improved two-stage mixed language model approach for handling out-of-vocabulary words in large vocabulary continuous speech recognition,” *Computer Speech and Language*, vol. 28, pp. 141-162, 2014.

[7] H. Bourlard, B. D'hoore and J. M. Boite, “Optimizing recognition and rejection performance in wordspotting systems,” in *Proc. 1994 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. I-373-I-376.

[8] Z. Junfeng and Z. Yeping, “A multi-confidence feature combination rejection method for robust speech recognition,” in *Proc. 2011 IEEE Int. Conf. Transportation, Mechanical, and Electrical Engineering*, pp. 2556-2559.

[9] J. Veth, B. Cranen, and L. Boves, “Acoustic backing-off as an implementation of missing feature theory,” *Speech communication*, vol. 34, pp. 247-256, 2001.

[10] J. Zhang, S. Kwong, G. Wei and Q. Y. Hong, “Using Mel-frequency cepstral coefficients in missing data technique,” *EURASIP Journal on Applied Signal Processing*, vol. 3, pp. 340-346, 2004.

[11] H. Christensen, B. Lindberg, and O. Andersen, “Employing heterogeneous information in a multi-stream framework,” in *Proc. 2000 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1571-1574.

[12] S. Young, and etc. (2006, December). The HTK Book (for HTK version 3.4) [Online]. Available: http://htk.eng.cam.ac.uk/