論文 / 著書情報
Article / Book Information

Title	VAD-measure-embedded Decoder with Online Model Adaptation
Authors	Tasuku Oonishi, Koji Iwano, Sadaoki Furui
Citation	INTERSPEECH2010, , , pp. 3122-3125,
Pub. date	2010, 9
Copyright	(c) 2010 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

VAD-measure-embedded Decoder with Online Model Adaptation

Tasuku Oonishi[1], *Koji Iwano*[2], *Sadaoki Furui*[1]

[1]Department of Computer Science, Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo, Japan
[2]Faculty of Environmental and Information Studies, Tokyo City University
3-3-1 Ushikubo-nishi, Tsuzuki-ku, Yokohama, Japan
oonishi@furui.cs.titech.ac.jp, iwano@tcu.ac.jp, furui@cs.titech.ac.jp

Abstract

We previously proposed a decoding method for automatic speech recognition utilizing hypothesis scores weighted by voice activity detection (VAD)-measures. This method uses two Gaussian mixture models (GMMs) to obtain confidence measures: one for speech, the other for non-speech. To achieve good search performance, we need to adapt the GMMs properly for input utterances and environmental noise. We describe a new unsupervised on-line GMM adaptation method based on MAP estimation. The robustness of our method is further improved by weighting updating parameters of GMMs according to the confidence measure for the adaptation data. We also describe an approach to accelerate the adaptation by caching statistical values to adapt GMMs. Experimental results on Drivers' Japanese Speech Corpus in a Car Environment (DJSC) show that the adaptation with decoding method significantly improves the word accuracy from 54.8% to 59.6%. Moreover, the weighting method improves the robustness of the unsupervised adaptation, and the cache method greatly accelerates the decoding process. Consequently, our adaptive decoding method significantly improves the word accuracy in a noisy environment with only a minor increase in the computational cost.

Index Terms: speech recognition, voice activity detection, Gaussian mixture model adaptation

1. Introduction

When a speech recognition system is deployed in a real environment, for example, to control a car navigation system or to use an automated telephone service, the recognizer often needs to process input signals that contain long non-speech or environmental noise periods. These non-speech periods must be correctly detected and removed, otherwise the decoder attempts to recognize the non-speech sounds as spoken utterances, and the recognition accuracy degrades because of insertion errors. Therefore, voice activity detection (VAD) is an essential part of a practical speech recognition system.

In a conventional VAD mechanism implemented at the front-end of a recognition system, speech and non-speech are classified at each input frame. If an input frame is detected as speech, then the frame is passed through the recognition pipeline. Otherwise, the frame is dropped at the front-end and not used for recognition. Typical VAD methods calculate a Speech/Non-Speech (SNS) score, and if the score exceeds some pre-determined threshold, the input frame is judged as speech. If the score is below the threshold, the input frame is judged as non-speech. Most common methods for measuring the SNS score utilize energy, the zero crossing rate (ZCR) [1], or the likelihood ratio. The likelihood ratio is calculated using Gaussian mixture models (GMMs) to model the speech and non-speech statistics [2].

Decisively classifying speech and non-speech in a front-end based VAD approach is often difficult. This is because calculating the SNS score precisely in noisy conditions is difficult. The decisive classification of speech and non-speech causes errors of discarding speech frames as non-speech frames, and vice versa.

To solve this problem, we previously proposed a novel search method making use of the SNS score on a frame-by-frame basis to bias hypothesis scores in the decoding phase [3]. If a hypothesis state belongs to silence or a short pause, a confidence measure as non-speech is multiplied to the frame acoustic score. Alternatively, if a hypothesis state belongs to a phoneme, a confidence measure is multiplied as speech. This method can reduce the number of errors of discarding speech frames as non-speech. Experiments using Drivers' Japanese Speech Corpus in a Car Environment (DJSC) [4] have shown that our method achieves 6 to 7% higher word accuracy than front-end VAD-based recognition systems. However, the recognition accuracy of our method was 7% lower than an oracle result obtained when correct speech/non-speech classification was given. This means that our search method still has room for improvement in performance.

One of the major causes of a deterioration in word accuracy in comparison with the oracle result is a mismatch between the speech and non-speech GMMs and actual speech and environmental signals. Therefore, acoustic model adaptation of GMMs should improve the word accuracy. Because measuring the actual input environment and adapting the GMMs before recognition in practical recognition systems are difficult, unsupervised on-line adaptation is essential. We investigated such an unsupervised adaptation approach for implementation in our search method.

Our adaptation approach uses the following two methods. The first is a parameter estimation method proposed by Reynolds [5] that uses MAP estimation. The second is an unsupervised adaptation method for GMMs proposed by Zhang [6]. We devised a method to improve the robustness of the adaptation in combination with these methods by weighting estimated GMM parameters based on confidence measures for adaptation data. We also devised an approach to accelerate the adaptation by caching statistical values of GMMs.

The rest of the paper is structured as follows; The next section describes details of our search method. In section 3 we describe our unsupervised on-line adaptation approach. In section 4 we demonstrate the effectiveness of our adaptation method

combined with our search method for an in-car speech recognition task. The paper finishes with conclusions and future work.

2. Search method utilizing SNS scores

In our search method using SNS scores [3], the acoustic likelihood of the i^{th} frame is biased by the confidence measure for speech or non-speech. If the hypothesis of the frame belongs to a phone model, the acoustic likelihood is biased by equation (1). Otherwise, the hypothesis belongs to a silence model, and the acoustic likelihood is biased by equation (3).

$$\log \hat{p}_{am}(X_i|\theta_v) = \log p_{am}(X_i|\theta_v) + \alpha \log \bar{C}^i_{H_1} \quad (1)$$

$$\bar{C}^i_{H_1} = \frac{\sum_{k=i-l}^{i+l} p(X_k|H_1)}{\sum_{k=i-l}^{i+l}\{p(X_k|H_0) + p(X_k|H_1)\}} \quad (2)$$

$$\log \hat{p}_{am}(X_i|\theta_{uv}) = \log p_{am}(X_i|\theta_{uv}) + \alpha \log \bar{C}^i_{H_0} \quad (3)$$

$$\bar{C}^i_{H_0} = \frac{\sum_{k=i-l}^{i+l} p(X_k|H_0)}{\sum_{k=i-l}^{i+l}\{p(X_k|H_0) + p(X_k|H_1)\}} \quad (4)$$

Here, X_i is the i^{th} feature vector, θ_v is a hypothesis of speech, θ_{uv} is a hypothesis of non-speech, $p_{am}(X_i|\theta_v)$ and $p_{am}(X_i|\theta_{uv})$ are acoustic model scores, α is a scaling factor, and l is a smoothing parameter for computing $\bar{C}^i_{H_1}$ and $\bar{C}^i_{H_0}$ over a window. $\bar{C}^i_{H_1}$ is a smoothed confidence measure of speech, and $\bar{C}^i_{H_0}$ is a smoothed confidence measure of non-speech. The confidence measures are normalized between 0 and 1. $p(X_i|H_1)$, and $p(X_i|H_0)$ are calculated using speech and non-speech GMMs.

The smoothed confidence measures for speech and non-speech are calculated using the likelihood value with GMMs averaged over $i-l$ to $i+l$ frames. They can also be calculated by averaging confidence measures over $i-l$ to $i+l$ frames [3]. Because a preliminary experiment showed better word accuracy by smoothing the likelihood of GMMs, we decided to use this likelihood smoothing in this study.

3. On-line unsupervised GMM adaptation

3.1. Adaptation process

Let X be a D dimensional feature vector for a frame. A likelihood of X given a GMM is calculated by:

$$p(X|\lambda) = \sum_{m=1}^{M} w_m p_m(X), \quad (5)$$

where λ is a parameter set of GMM, M is the number of mixtures, and $p_m(X)$ is a probability density function (PDF) of the m^{th} Gaussian component. w_m is a mixture weight and $\sum_m w_m$ equals 1. The PDF is defined by a D dimensional normal distribution:

$$p_m(X) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_m|^{\frac{1}{2}}} \cdot \exp\{-\frac{1}{2}(X-\mu_m)^\top \Sigma_m^{-1}(X-\mu_m)\}, \quad (6)$$

where μ_m is the D dimensional mean vector, and Σ_m is a $D \times D$ variance-covariance matrix. Therefore, the model parameter set is $\lambda : \{\mu_m, \Sigma_m, w_m\}(m = 1, 2, \cdots, M)$.

Reynolds proposed an adaptive training method for GMMs based on the MAP estimation [5]. In this method, parameters of a GMM are estimated using the following two steps.

Step1: Calculate n_m, $E_m(X)$, and $E_m(X^2)$ with an adaptation data sample set, $X_1, X_2, \cdots, X_N$

$$Pr(m|X_i) = \frac{w_m p_m(X_i)}{\sum_{k=1}^{M} w_k p_k(X_i)} \quad (7)$$

$$n_m = \sum_{i=1}^{N} Pr(m|X_i) \quad (8)$$

$$E_m(X) = \frac{1}{n_m}\sum_{i=1}^{N} Pr(m|X_i)X_i \quad (9)$$

$$E_m(X^2) = \frac{1}{n_m}\sum_{i=1}^{N} X_i^\top Pr(m|X_i)X_i \quad (10)$$

$Pr(m|X_i)$ is a responsibility, and $\sum_m Pr(m|X)$ equals 1.

Step2: Update the parameters of the GMM

$$\hat{w}_m = [\beta_m n_m/N + (1-\beta_m)w_m]\rho \quad (11)$$

$$\hat{\mu}_m = \beta_m E_m(X) + (1-\beta_m)\mu_m \quad (12)$$

$$\hat{\Sigma}_m = \beta_m E_m(X^2) + (1-\beta_m)(\Sigma_m + \mu_m\mu_m^\top) - \hat{\mu}_m\hat{\mu}_m^\top \quad (13)$$

$\hat{w}_m, \hat{\mu}_m$, and $\hat{\Sigma}_m$ are updated parameters of the weight, the mean, and the variance-covariance matrix for the m^{th} Gaussian. ρ is computed for all the adapted mixture weights to ensure they sum up to 1. w_m, μ_m, and Σ_m are the weight, the mean, and the variance-covariance matrix for the m^{th} Gaussian before adaptation. β_m controls the relative balance between the old and new statistics:

$$\beta_m = \frac{n_m}{n_m + \gamma}, \quad (14)$$

where γ is an application dependent parameter. If γ is set to 0, the updated parameters equal the new statistics, and if γ is set to ∞, the updated parameters equal the old statistics.

In our on-line adaptation process for GMMs, GMM parameters at the $i+1^{\text{th}}$ frame are updated using an adaptation data sample set $X_1, X_2 \cdots X_i$. Confidence measures for speech and non-speech at the $i+1^{\text{th}}$ frame are calculated with updated GMM parameters. At the $i+2^{\text{th}}$ frame, the GMM parameters are updated again using an adaptation data sample set $X_1, X_2 \cdots X_{i+1}$ and the initial GMM. This process is repeated until the end of the input signal.

3.2. Adaptation data selection

To properly update the GMM parameters, we utilize a similar strategy as Zhang [6] to select adaptation data by setting the threshold according to the confidence measure. If the confidence measure for speech X_i at the i^{th} frame exceeds threshold τ, X_i is added to an adaptation data set for the speech model. If the confidence measure for non-speech at the i^{th} frame exceeds the threshold, X_i is added to an adaptation data set for the non-speech model. If neither the confidence measure for speech nor non-speech exceeds the threshold, X_i is discarded. Finally, if confidence measures for both speech and non-speech exceed the threshold, X_i is added to both of the adaptation data sets.

It is difficult to optimize the τ parameter, which controls the unsupervised adaptation data selection, according to actual

acoustic conditions, such as noise conditions and original parameters of GMMs. Therefore, the robustness of the adaptation method against mismatched τ parameter must be achieved so that it can be easily fixed before using the recognition system.

For this purpose, we utilize a method to adjust parameter β_m, which controls the balance between old and new statistics according to a confidence measure for the adaptation data set. An estimated parameter using an adaptation data set whose confidence measures are relatively high is assumed to be more accurate than an estimated parameter using an adaptation data set whose confidence measures are relatively low. To implement this assumption, we weight the responsibility $Pr(m|X_i)$ for the m^{th} Gaussian using a confidence measure of the i^{th} frame. The weighted responsibility $\hat{Pr}(m|X_i)$ is used to estimate the parameters instead of using $Pr(m|X_i)$.

$$\hat{Pr}(m|X_i) \quad = \quad C^i_{H_j} \times Pr(m|X_i),\ j \in 0 \text{ or } 1 \quad (15)$$

This method adjusts β_m based on the confidence measure of an adaptation data set. If all confidence measures of an adaptation data set are 0, n_m equals 0, which means that β_m equals 0 and the GMM parameters are not updated.

3.3. Fast adaptation

Because the on-line adaptation approach described in Section 3.1 recalculates statistical values of GMMs at each frame, a high computational cost is incurred. However, our on-line adaptation approach calculates the statistical values, $n_m, E_m(X)$ and $E_m(X^2)$, using the initial model instead of using the updated model as the Zhang's method [6]. Our method allows caching and reusing the statistical values of previous frames, thereby reducing the cost and accelerating the adaptation speed. In our method, the statistical values, n_m, $Pr(m|X_i)X_i$, and $X_i^\top Pr(m|X_i)X_i$, are cached at each frame. These values are then reused to calculate $n_m, E_m(X)$, and $E_m(X^2)$ in the **Step1** described in Section 3.1.

4. Experiments

We evaluated our methods using the Drivers' Japanese Speech Corpus in a Car Environment (DJSC) [4]. This consisted of utterances in a hands-free command-and-control task recorded in a car driven on a motorway. The test set consisted of utterances by 40 speakers equally distributed between male and female speakers. The S/N ratio of the test set was varied between -8dB and 0dB [4]. Each speaker provided 41 command utterances spoken with a style used in operating a car navigation device while driving. The command utterances were separated by non-speech periods for one to two seconds, including various background noise. The utterances were recorded using a microphone mounted on the navigation device and were sampled at 16 kHz.

The acoustic models were trained on 52 hours of speech data from the Japanese Newspaper Article Sentences (JNAS [7]) corpus. The training material was gender balanced, containing 130 male speakers giving 25 hours of speech and 130 female speakers providing another 27 hours of speech.

The training and testing data were processed during the evaluation as follows. Raw speech waveforms were converted to a sequence of 38 dimensional feature vectors with a 10ms frame rate and 25ms window size. Each feature vector was composed of 12 Mel-frequency cepstral coefficients (MFCCs) with deltas and delta-deltas, augmented with log delta and delta-delta energy terms. The acoustic models were EM trained using

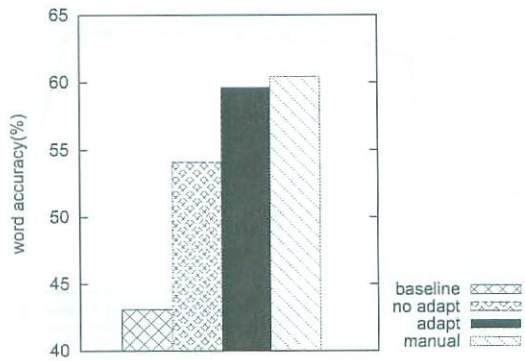

Figure 1: *Effect of on-line unsupervised adaptation*

the processed data, and this process yielded a set of three states, left-to-right tri-phone HMMs with 2000 states. Each state output density was represented by a 16 component GMM, with each component having a diagonal covariance.

The language model was a network grammar, and the vocabulary was 83 words covering all commands. The network had a path that corresponded to each of the valid commands that looped through the initial state to allow continuous recognition of an utterance stream.

The GMMs to model speech and non-speech had four Gaussian components. The speech GMM was trained using the data from 967 lectures in the Corpus of Spontaneous Japanese (CSJ) [8], and the non-speech GMM was trained with data from car noise in the Japan Electronic Industry Development Association (JEIDA) database. Unsupervised on-line adaptation was performed independently for each speaker. The utterance length including silence was 150 to 200 seconds for each speaker. In the recognition evaluation, we used the T^3 Decoder [9] developed at the Tokyo Institute of Technology. The experiments were conducted on 2.5-GHz Intel Xeon machines. The scaling factor α was fixed to 3, and the smoothing parameter l was fixed to 15 throughout the experiments.

4.1. Effects of unsupervised on-line adaptation

Figure 1 shows the recognition accuracy when using the unsupervised on-line adaptation:

- "baseline" represents the result without implementing VAD.
- "no adapt" represents the result of our search method with no GMM adaptation.
- "adapt" represents the result with unsupervised on-line adaptation in our search method.
- "manual" corresponds to the result when using the corpus manually labeled to remove all non-speech periods.

In these experiments, γ and τ parameters were fixed to 10 and 0.7, respectively, based on the results of our preliminary experiments. The results in the figure show that the word recognition accuracy was 43.1% and 54.8% without and with the VAD. The unsupervised on-line adaptation further improved the accuracy to 59.6%, meaning that it was 4.8% higher in absolute value than the accuracy without adaptation. The result was close to the oracle word accuracy: 60.4%. These experimental results show that the adaptation approach implemented in our search method can significantly improve word accuracy, and it can achieve almost the same result as the ideal segmentation case.

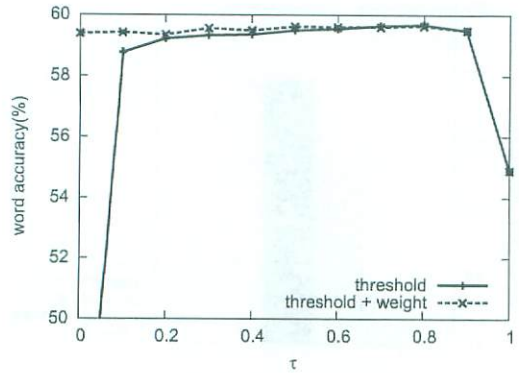

Figure 2: *Robustness against τ parameter*

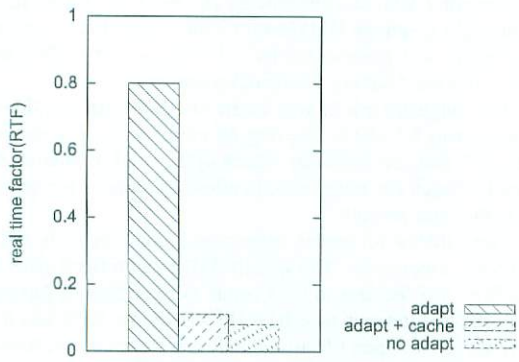

Figure 3: *Effect of caching for fast adaptation*

4.2. Robustness of the unsupervised adaptation

Figure 2 shows the effectiveness of weighting β_m based on the confidence measure obtained for an adaptation data set. The horizontal and vertical axes correspond to τ and word accuracy. The parameter γ was set to 10. "threshold" in the figure shows the results without weighting, and "threshold + weight" shows that with weighted β_m. The former result shows that the word accuracy degrades when the τ value decrease. However, the latter result shows that the robustness of our method significantly increases against the variation in the τ value, and significant loss in word accuracy can be avoided by weighting β_m.

4.3. Effects of fast adaptation

Figure 3 shows the effectiveness of the fast adaptation method in terms of the real time factor (RTF). The parameters γ and τ were fixed at 10 and 0.7, respectively. "adapt" in the figure shows the result for the unsupervised on-line adaptation without caching, "adapt + cache" shows the result with the caching described in Section 3.3, and "no adapt" shows the result without adaptation. This figure shows that the recognition time significantly drops by using the caching, and it becomes almost equivalent to the result with no adaptation, meaning that most of the computational cost for adaptation was eliminated. These results show the effectiveness of the caching in our on-line unsupervised adaptation.

5. Conclusion

This paper presented a new approach for adapting GMMs to an input environment in our search method, in which recognition hypotheses are biased by VAD values. Experimental results on the DJSC database show that our unsupervised on-line adaptation method can significantly improve word accuracy in a noisy environment. Moreover, the weighting method using a confidence measure of an adaptation data set can improve the robustness of the system against variation in the threshold parameter for selecting the data for adaptation, and the caching method can significantly accelerate the adaptation speed.

In future work, we will conduct further evaluations on the robustness of the adaptation method in various noisy and SNR conditions, and we will devise an automatic method to optimize the γ parameter according to the task and environmental conditions.

6. Acknowledgements

The Drivers' Japanese Speech Corpus in a Car Environment was recorded by Asahi Kasei Corp. under the Development of Fundamental Speech Recognition Technology project supported by the Japanese Ministry of Economy, Trade and Industry. We would like to thank Asahi Kasei Corp. for letting us use the corpus.

7. References

[1] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin and J.-P Petit, "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," IEEE Communications Magazine 35 (9): pp. 64-73, 1997.

[2] R. Singh, M. Seltzer, B. Raj and M. Stern, "Speech in Noisy Environments: robust automatic segmentation, feature extraction, and hypothesis combination," Proc. ICASSP, vol. 1, pp. 273-276, 2001.

[3] T. Oonishi, P. Dixon, K. Iwano and S. Furui, "Robust speech recognition using VAD-measure-embedded decoder," Proc. Interspeech, pp. 2239-2242, 2009.

[4] K. Hiraki, T. Shinozaki, K. Iwano, A. Betkowska, K. Shinoda and S. Furui, "Initial evaluation of the driver's Japanese speech corpus in a car environment," IEICE Technical Reports, Asian Workshop on speech science and Technology, SP-2007-202, pp. 93-98, 2008.

[5] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, 2000.

[6] Y. Zhang and M. Scordilis, "Effective online unsupervised adaptation of Gaussian mixture models and its application to speech classification," Pattern Recognition Letters, vol. 29, no. 6, pp. 735-744, 2008.

[7] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," J.Acoust. Soc. Jpn.(E), vol. 20, no. 3, pp. 199-206, 1999.

[8] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. 7-12, 2003.

[9] P. Dixon, D. Caseiro, T. Oonishi and S. Furui, "The TITECH large vocabulary WFST speech recognition system," Proc. IEEE Workshop on ASRU, pp. 443-448, 2007.