Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Speech recognition in noise by using word graph combinations

Shunsuke Kuramata, Masaharu Kato and Tetsuo Kosaka

Graduate School of Science and Engineering, Yamagata University, 4-3-16 Jonan Yonezawa-City, Yamagata, Japan

PACS: 43.72.Ne

ABSTRACT

In the practice, the performance of speech recognition systems is affected by speech signals being corrupted with various background noises in the environment. In this paper, we propose a new word graph combination (WGC) approach for speech-in-noise recognition. The aim of this work is to develop a method that would ensure robust speech recognition under various noise conditions, and in particular, under the adverse effect of environmental and impulsive noise. For this purpose, we developed a word graph combination (WGC) technique in which both continuous-mixture hidden Markov models (CMHMMs) and discrete-mixture hidden Markov models (DMHMMs) are being used as acoustic models. It has been previously verified that a DMHMM-based system can ensure significant improvements in the speech recognition performance under impulsive noise conditions. We also showed that the CMHMM-based system indicated better performance in high SNR conditions and environmental noise conditions. On the grounds of the above mentioned findings, we adopted a system combination approach in which both a DMHMM and a CMHMM are used. With the proposed method, complementary effects can be anticipated because the CMHMM and the DMHMM exhibit different error trends. Among the existing combination methods, which include recognizer output voting for error reduction (ROVER) and confusion network combination (CNC), in our work, we selected the technique of WGC. Unlike conventional combination approaches, like ROVER and CNC, the timing information for all word hypotheses is well preserved in the WGC. In the speech recognition experiments we performed, the proposed system showed better performance than the ROVER-based system or the baseline system. In particular, this new system showed comparatively higher performance under mixed noise conditions.

INTRODUCTION

In this study, we aim to improve speech recognition under noisy conditions by using the word graph combination (WGC) technique [1]. In the past studies, a multitude of techniques has been proposed, all of which are based on a combination of systems exhibiting different error trends. These include recognizer output voting for error reduction (ROVER) [2] and confusion network combination (CNC) [3], which are both popular techniques. The system combination technique was originally proposed as a means of combining systems from multiple sites [4]. More specifically, the aim of this combination was to improve speech recognition performance by combining the results from individual systems developed by different organizations. In contrast, we tried to improve the speech recognition performance by combining several acoustic models which are all characterized by different error trends.

Some of our previous work was focused on improving the performance of speech-in-noise recognition through the combination of outputs from continuous-mixture hidden Markov model (CMHMM) and discrete-mixture hidden Markov model (DMHMM) by ROVER [5]. The previous work used the differences of the error trends in the results of CMHMM and DMHMM. DMHMM uses a mixture of discrete distributions which have a high degree of flexibility, and are expected to represent complicated shapes such as noisy conditions. As a consequence, under impulsive noise conditions, DMHMM shows a higher recognition rate than CMHMM [6]. Moreover, speech recognition error trends differ between CMHMM and DMHMM because of differences in the acoustic conditions, for example the signal-to-noise ratio (SNR) and the type of noise. Therefore, this new approach of using both CMHMM and DMHMM as acoustic models can improve the speech recognition performance in various noise environments.

This work consists of combining the CMHMM and the DMHMM using the WGC technique. The difference between ROVER and WGC lies in the combination of the speech recognition results, or the combination of the intermediate results. On the other hand, an advantage of WGC over ROVER is that it supports rescoring on the integrated hypothesis space. Furthermore, the timing of words is kept explicit in the structure of the word graph. Therefore, WGC can expect to provide for a significantly improved speech recognition performance when compared to the ROVER system.

In order to verify the effectiveness of the proposed method, we performed speech recognition experiments under various noise conditions. In the experiments, we used three types of noises: environmental noise signals which were either stationary or slow-varying, impulsive noise in which the power and spectral features may radically change within a very short time, and mixed noise where the environmental noise and the impulsive noise were mixed artificially.

ACOUSTIC MODELS

In this paper, CMHMM and DMHMM are used as acoustic models for system combination. This section provides a detailed description of these models.

Continuous-mixture hidden Markov models (CMHMMs)

In speech recognition, CMHMMs have been widely used as acoustic models, in which the output probability density is modeled by a mixture of Gaussian distributions as follows:

$$b_i(\boldsymbol{o}_t) = \sum_m w_{im} N\left(\boldsymbol{o}_t \middle| \mu_m, \Sigma_m\right). \tag{1}$$

Here, $N(\boldsymbol{o} \mid \mu_m, \Sigma_m)$ is a Gaussian distribution with mean μ_m and covariance Σ_m, while w_{im} is the mixture weight of the m[th] distribution. In this paper, the diagonal covariance was used in consideration of computational cost.

Discrete-mixture hidden Markov models (DMHMMs)

The DMHMM is a type of discrete hidden Markov model (DHMM) that was originally proposed by Takahashi *et al.* to reduce computational cost in decoding processes [7]. More specifically, two types of the DMHMMs have been proposed. In the first scalar-based quantization is being used [7] and subvector-based quantization in the second [8]. In this paper, we have employed subvector-based DMHMMs. In the subvector-based method, the feature vector is partitioned into S subvectors, $\boldsymbol{o}_t = [\boldsymbol{o}_{1t}, \ldots, \boldsymbol{o}_{st}, \ldots, \boldsymbol{o}_{St}]$ and vector quantization (VQ) codebooks are provided for each subvector. Subsequently, the feature vector $\boldsymbol{o}_t$ is quantized as follows:

$$q(\boldsymbol{o}_t) = [q_1(\boldsymbol{o}_{1t}), \ldots, q_s(\boldsymbol{o}_{st}), \ldots, q_S(\boldsymbol{o}_{St})], \tag{2}$$

where $q_s(\boldsymbol{o}_{st})$ is the discrete symbol for the s[th] subvector. The output distribution of the DMHMM, $b_i(\boldsymbol{o}_t)$, is given by the expression:

$$b_i(\boldsymbol{o}_t) = \sum_m w_{im} \prod_s \hat{p}_{sim}(q_s(\boldsymbol{o}_{St})), \tag{3}$$

where w_{im} is the mixture coefficient for the m[th] mixture in state i, and $\hat{p}_{sim}(q_s(\boldsymbol{o}_{St}))$ is the probability of the discrete symbol for the s[th] subvector.

In the remainder of this section, we describe the method used for the estimation of the DMHMM parameters based on the maximum *a posteriori* (MAP) criterion. An estimate of the maximum likelihood (ML) of the discrete probability $p_{sim}(k)$ is calculated with the use of the following expression:

$$p_{sim}(k) = \frac{\sum_{t=1}^{T} \gamma_{imt}\, \delta(q_s(\boldsymbol{o}_{st}), k)}{\sum_{t=1}^{T} \gamma_{imt}}, \tag{4}$$

where k is the index of the subvector codebook and γ_{imt} is the probability of the m[th] mixture component being in state i at time t. If we assume that the prior distribution is represented by the Dirichlet distribution, the estimate of the DMHMM $\hat{p}_{sim}(k)$ based on the MAP criterion is given by the following expression:

$$\hat{p}_{sim}(k) = \frac{\tau \cdot p^0_{sim}(k) + n_{im} \cdot p_{sim}(k)}{\tau + n_{im}}, \tag{5}$$

where $p^0_{sim}(k)$ is the constrained prior value of the discrete probability and τ indicates the relative balance between the corresponding prior value and the observed data. In our experiments, τ was set to 10.0 based on the results of comparative experiments. Although both the mixture coefficient and transition probability can be estimated by means of the MAP criterion, only the output probability is being estimated in this paper.

Compensation of DMHMMs

To improve the noise robustness of speech recognition, a compensation method for discrete distributions is applied. It is more likely that a significant reduction of the output probability will appear under severe mismatch conditions caused by unknown noise. This method can alleviate the adverse effect of the unknown noise during the decoding process. If one of the subvector probabilities, $\hat{p}_{sim}(q_s(\boldsymbol{o}_{st}))$, in Eq. (3) is close to 0, the output probability, $b_i(\boldsymbol{o}_t)$, will also be close to 0. In this case, noise will have a detrimental effect on the decoding process, even if the time length of exposure to noise is short. In the compensation of DMHMMs, a threshold is set for the discrete probability, and the detrimental effect of noise is reduced. The compensation method can be described as follows: if in Eq. (3) $\hat{p}_{sim}(q_s(\boldsymbol{o}_{st})) < dth$, the output probability is set to *dth*, where *dth* is the threshold value for the subvector. The threshold was set to 0.00025 in this paper.

WORD GRAPH COMBINATION (WGC)

Overview

In this section, we describe the system combination approach proposed for speech-in-noise recognition. It aims to raise the speech recognition rate by combining the output of CMHMM with that of DMHMM. The system combination approach has been proven to result in significant improvements if the speech recognition results are substantially different between systems. Since the speech recognition results from DMHMMs and CMHMMs are different between each other under noisy conditions, the performance is expected to improve. The procedure followed for WGC is provided below.

1. An input speech is decoded using two acoustic models (AM1, AM2), and a bigram language model. Then, two word graphs, WG1 and WG2, are obtained by the decoding process.
2. The two word graphs (WG1, WG2) are combined to form one single word graph, WG_C.
3. The word graph WG_C is rescored using the two acoustic models and a trigram language model. In this step, the two scores obtained by AM1 and AM2 are merged to obtain one single score. Further details of the scoring process are available in the next section.

The procedure of the WGC is shown in Figure 1.

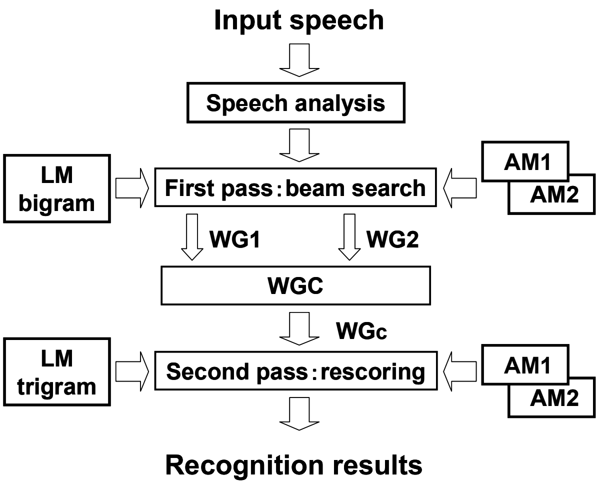

Figure 1. Block diagram of the WGC technique.

Algorithm of WGC

Suppose that there are N word graphs, $W_1, W_2, ..., W_N$, to be combined. If two arcs, q_1 in W_1 and q_2 in W_2, are equal, the two word graphs, W_1 and W_2, can be combined as follows:

$$W_1 + W_2 = \{q = q_1 + q_2 | q_1 = q_2\} \cup \{q_1 | q_1 \notin W_2\} \cup \{q_2 | q_2 \notin W_1\}. \tag{6}$$

Two equal arcs are defined as equal when they have the same word ID, start time and end time. A detailed description of the algorithm behind WGC has been provided by Chen and Lee [1].

Scoring methods

In this section, two types of scoring methods are described. They are denoted by the terms "average score" and "weighted score". According to the "average score" method, the merged score P'_m for the m^{th} edge is given by the expression:

$$P'_m = \frac{1}{N} \sum_{k=1}^{N} p_m^k, \tag{7}$$

where N is the number of systems, and p_m^k is the score for the m^{th} edge in system k. In this study, N has been set to two. The "weighted score" of m^{th} edge is calculated as follows:

$$P'_m = (1-\alpha) P_m^1 + \alpha P_m^2, \tag{8}$$

where P_m^1 is the score calculated by CMHMMs, and P_m^2 is the score provided by DMHMMs, while α indicates the balance between CMHMMs and DMHMMs.

EXPERIMENTAL SETUP

We used the "Japanese Newspaper Article Sentences" (JNAS) as training and test data. More specifically, two sets of training data were used; one for clean training, and the other for multi-condition training [9]. The training data set consists of 15,732 Japanese sentences uttered by 102 male speakers. For clean training, no noise had been added to the data. On the other hand, for multi-condition training, all these utterances were divided into 20 subsets. No noise was added to four subsets. In the remainder of the data, noise was artificially added. Four types of noise–train, crowd, car and exhibition hall–were selected and added to the utterances at a SNR of 20, 15, 10 and 5 dB. We used four types of acoustic models; clean condition CMHMMs (CC CMHMM), clean condition DMHMMs (CC DMHMM), multi-condition CMHMMs (MC CMHMM) and multi-condition DMHMMs (MC DMHMM) in our speech recognition experiments. The training method employed for each one of these acoustic models is explained below. The CC CMHMMs were trained by ML estimation using the clean training data. The CC DMHMMs were trained by MAP estimation using clean training data. The initial models used for their training were derived from the conversion of the CC CMHMMs into DMHMMs. The MC CMHMMs were trained by ML estimation using multi-condition training data and CC CMHMMs were used as initial models. The MC DMHMMs were trained by MAP estimation using multi-condition training data. In this case, the initial models were obtained by converting MC CMHMMs into DMHMMs. The three types of test sets used were as follows.

- Test set for environmental noise
  Four types of noise–station, factory, street crossing and elevator hall–were added to 100 sentences uttered by 10 male speakers at a SNR of 20, 15, 10 and 5 dB. These noises were different from those of the training data.
- Test set for impulsive noise
  The impulsive noise signals were added to 100 sentences uttered by 10 male speakers. Three types of impulsive noise were selected from the Real World Computing Partnership (RWCP) database [10], namely:
  whistle3 blowing a whistle;
  claps1 handclaps;
  bank hitting a coin bank.
  These noise signals were added to speech data at intervals of 1 sec and a SNR of 0 dB. The SNR was calculated as the average power of the speech data divided by the maximum power of the impulsive noise. The maximum power was determined from power values that were calculated from the impulsive noise data every 30 msec.
- Test set for mixed noise
  Noise signals from above two test sets were mixed artificially to make a new test set. Four types of noises were prepared at a SNR of 10 dB as environmental noises. These noises were mixed with three impulsive noises. Thus, twelve types of noises were used for evaluation of the proposed system.

The speech analysis conditions are summarized in Table 1. The structure of CMHMM and DMHMM was 2000-state HMnet (set of shared state triphones), and the number of mixture components was 16. Table 2 summarizes the subvector allocation and the codebook size for DMHMM. Although Δ and Δ^2 have been omitted from the table, all the codebooks were designed in the same manner. A two-pass search decoder using a bigram and a trigram was used in speech recognition. Decoding was performed in the first pass by means of a one-pass algorithm, in which a frame-synchronous beam search algorithm and a tree-structured lexicon were applied. The bigram and trigram models were trained using 45 months of newspapers article sentences. The trained language models had 5 K word entries.

In the case of ROVER, 50 different outputs are combined. These 50 outputs were obtained by varying parameters such as language weight and word insertion penalty. More specifi-

Table 1. Speech analysis conditions

Sampling frequency	16 kHz
Quantization	16 bit
Frame length	32 msec
Frame period	8 msec
Analysis window	Hamming window
Feature vector	MFCC (1-12), log power+ $\Delta + \Delta\Delta$ (total of 39 dimensions)
Normalization	CMN

Table 2. Codebook size for each subvector

Parameter	logP	C_1, C_2	C_3, C_4	C_5, C_6	C_7, C_8	C_9, C_{10}	C_{11}, C_{12}
CB size	64	64	64	64	64	64	64

Table 3. Values of language weight and word insertion penalty for each acoustic model and noise condition

Environmental noise	CMHMM	language weight	12 ~ 24
		insertion penalty	-8 ~ -64
	DMHMM	language weight	12 ~ 24
		insertion penalty	-8 ~ -48
Impulsive noise	CMHMM	language weight	21 ~ 33
		insertion penalty	-48 ~ -80
	DMHMM	language weight	18 ~ 34
		insertion penalty	-40 ~ -72
Mixed noise	CMHMM	language weight	16 ~ 28
		insertion penalty	0 ~ -72
	DMHMM	language weight	14 ~ 30
		insertion penalty	-14 ~ -66

cally, five different values were used for the language weight and five values for the word insertion penalty. Namely, 5 × 5 = 25 types of parameter sets were prepared for both the CMHMMs and the DMHMMs. In the end, a total of 50 outputs were combined. The values of language weight and insertion penalty in each noise condition and acoustic model are summarized in Table 3. These parameters were set based on the results of prior experiments.

RESULTS AND DISCUSSION

Speech recognition experiments under environmental noise conditions

The experimental results of the combination of the MC CMHMM with the MC DMHMM under environmental noise conditions are summarized in Table 4. Here, the word error rate (WER) is provided for each model used. The term "structure combination" means that the structures of two word graphs were combined, but not the scores. On the other hand, "score combination" means that both the structures and scores were combined. The speech recognition results of ROVER are provided for comparison.

Table 4. WERs (%) under environmental noise conditions

	Before combination		Structure combination	
SNR (dB)	MC CMHMM	MC DMHMM	MC CMHMM	MC DMHMM
∞	6.21	5.69	5.80	5.38
20	7.79	8.10	7.87	7.87
15	11.52	11.59	11.49	11.34
10	24.43	23.50	22.59	21.64
5	52.23	49.79	50.67	48.50
Ave.	22.94	22.21	22.13	21.34

	Score combination		ROVER
	Without weighting	α =0.9	
∞	5.80	5.59	5.49
20	7.89	7.97	7.61
15	11.36	11.44	11.44
10	21.87	21.30	23.42
5	48.86	47.90	50.60
Ave.	21.51	21.18	22.22

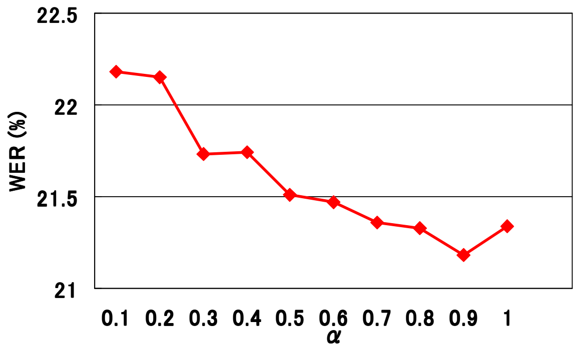

Figure 2. Relation between the parameter α and the WER (%) under environmental noise conditions

Based on the results, we concluded that both the structure and the score combination are effective. In the case of score combination, the performance achieved with weighting was slightly better than without weighting. The best performance can be obtained at α = 0.9 (see Figure 2). The reason why a large weight should be applied to the DMHMM is that the performance obtained with the use this model was better than CMHMM without combination. Moreover, the performance obtained with the proposed system was better than the ROVER system. On the ground of the above mentioned results, we concluded that the WGC approach is effective.

Table 5. WER (%) under impulsive noise conditions

	Before combination		Structure combination	
Noise	CC CMHMM	CC DMHMM	CC CMHMM	CC DMHMM
bank	10.46	8.49	8.59	8.28
claps1	12.53	9.21	12.73	10.04
whistle3	37.89	26.71	32.19	26.40
Ave.	20.29	14.80	17.84	14.91
	Score combination		ROVER	
	Without weighting	α =0.9		
bank	8.59	8.59	9.83	
claps1	11.08	9.63	11.39	
whistle3	27.12	26.19	34.89	
Ave.	15.60	14.80	18.70	

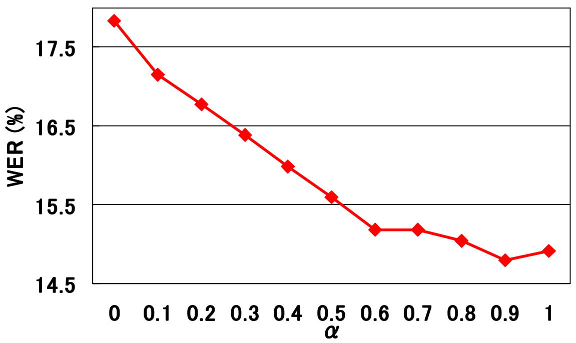

Figure 3. Relation between the parameter α and the WER (%) under impulsive noise conditions

Speech recognition experiments under impulsive noise conditions

The experimental results from the combination of the CC CMHMM with the CC DMHMM under impulsive noise conditions are summarized in Table 5. The relation between the parameter α and WER obtained with the score combination method is shown in Figure 3. We used clean condition models, because the duration of all the noise segments was very short and each utterance was almost entirely free of noise under impulsive noise conditions.

Based on the experimental results, we concluded that with the CC DMHMM before combination and the CC DMHMM after the score combination we can obtain approximately the same WER value (14.80%). Under impulsive noise conditions, the WGC method did not result in any improvement in the speech recognition performance. This was attributed to the difference in performance obtained with CMHMM and that of DMHMM before combination. If the system could have known in advance that the noise is impulsive, it would be recommended to use CC DMHMMs. However, the system generally can not predict the noise conditions. Therefore, the use of the WGC approach is suitable even if the speech recognition performance does not improve under impulsive noise conditions.

Table 6. WER (%) under mixed noise conditions

	Before combination		Structure combination	
Noise	MC CMHMM	MC DMHMM	MC CMHMM	MC DMHMM
bank	31.42	30.38	30.56	28.91
claps1	34.14	32.92	33.90	31.37
whistle3	52.56	54.19	50.88	50.52
Ave.	39.37	39.16	38.45	36.93
	Score combination		ROVER	
	Without weighting	α =0.8		
bank	29.19	28.73	29.94	
claps1	32.14	31.55	32.74	
whistle3	49.51	50.16	51.71	
Ave.	36.95	36.81	38.13	

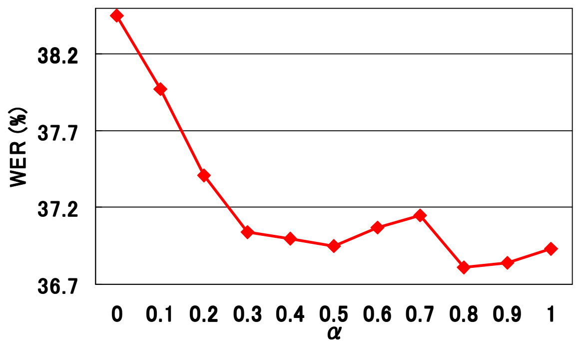

Figure 4. Relation between the parameter α and the WER (%) under mixed noise conditions

Speech recognition experiments under mixed noise conditions

The experimental results from the combination of the MC CMHMM with the MC DMHMM are summarized in Table 6. The relation between the parameter α and WER is shown in Figure 4. Each result represents an average WER of the four mixed noises combined every time with one of the impulsive noises.

Under mixed noise conditions, both the structure combination and the score combination are effective as well as environmental noise conditions described in the previous section. By comparison, the WGC approach was more effective under mixed noise conditions than under environmental noise conditions. Based on these findings, we concluded that the WGC technique is effective in various noise environments.

SUMMARY OF SPEECH RECOGNITION EXPERIMENTS

The experimental results for the three different speech recognition modes (before combination, structure combination and score combination) are summarized in Table 7. Based on these results, we concluded that both the structure combination and the score combination are effective. In particular, the

Table 7. Summary of WERs (%) for three recognition modes

Before combination		
Noise	CMHMM	DMHMM
Environmental noise	22.94	22.21
Impulsive noise	20.29	14.80
Mixed noise	39.37	39.16
Structure combination		
Environmental noise	22.13	21.34
Impulsive noise	17.84	14.91
Mixed noise	38.45	36.93
Score combination		
	Without weighting	With weighting
Environmental noise	21.51	21.18 (α =0.9)
Impulsive noise	15.60	14.80 (α =0.9)
Mixed noise	36.95	36.81 (α =0.8)

best possible performance could be achieved by using the score combination with the score weighting method.

CONCLUSIONS

In this paper, we propose a WGC method in which acoustic models that exhibit different error trends can be combined in order to improve the performance of speech-in-noise recognition. Based on experimental results obtained under environmental noise conditions and mixed noise conditions, we concluded that the WGC method can ensure significant improvements in the performance of speech recognition. More specifically, better speech recognition performance could be obtained with the proposed combination method than with ROVER, the conventional combination method. While our experimental results suggest that the WGC approach is effective under various noise conditions, the combination of weighted scores is also effective. The next step in this line of research would be to combine the ROVER with the WGC method in order to achieve further improvements. In addition, we have started working on the automatic estimation of the weighting coefficients used in the score combination.

REFERENCES

1 I-F. Chen and L.-S. Lee, “A new framework for system combination based on integrated hypothesis space” In Proc. of Interspeech-2006, pp. 533-536 (2006)

2 J. G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)” In Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 347-354 (1997)

3 L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks” Computer Speech and Language, **14**, 373-400 (2000)

4 M.J.F. Gales, D.Y. Kim, P.C. Woodland, H.Y. Chan, D. Mrva, R. Sinha and S.E. Tranter, “Progress in the CU-HTK broadcast news transcription system” IEEE Trans. on Audio, Speech and Language Processing, **14**, 1513-1525 (2006)

5 T. Kosaka, Y. Saito and M. Kato, “Noisy speech recognition by using output combination of discrete-mixture HMMs and continuous-mixture HMMs” In Proc. of Interspeech 2009, pp. 2379-2382 (2009)

6 T. Kosaka, M. Katoh and M. Kohda, “Robust speech recognition using discrete-mixture HMMs” IEICE–Transactions on Information and Systems, **E88-D**, 2811-2818 (2005)

7 S. Takahashi, K. Aikawa, and S. Sagayama: “Discrete mixture HMM” In Proc. of ICASSP97, pp. 971-974 (1997)

8 S. Tsakalidis, V. Digalakis, and L. Newmeyer, “Efficient speech recognition using subvector quantization and discrete-mixture HMMs” In Proc. of ICASSP99, pp. 569-572 (1999)

9 D. Pearce and H.-G. Hirsch: “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions” In Proc. of ICSLP2000, vol.4, pp. 29-32 (2000)

10 S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T.Yamada: “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition” In Proc. of ICLRE, pp. 965-968 (2000)