Vector-Quantization-Based Speech Recognition and Speaker Recognition Techniques

Sadaoki Furui
NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan

Abstract

This paper reviews major methods of applying the vector quantization (VQ) technique to speech and speaker recognition. These include speech recognition based on the combination of VQ and the DTW/HMM technique, VQ-distortion-based recognition, learning VQ algorithms, speaker adaptation by VQ-codebook mapping, and VQ-distortion-based speaker recognition. Not only has it reduced the amount of computation and storage, the VQ technique has also created new ideas of solving various problems in speech/speaker recognition.

1. Introduction

The vector quantization (VQ) technique was first applied to speech coding and image coding. In speech/speaker recognition, the VQ technique was also first used as an efficient spectral quantization technique, and it has greatly helped to reduce the amount of computation and storage. The VQ technique has also been used as a non-parametric representation method of spectral distribution, and this approach has created various new algorithms for speech/speaker recognition. The following chapters give an overview of these techniques.

2. VQ-Based Speech Recognition

2.1 Combination of VQ and DTW

Speech recognition is essentially based on the comparison between an utterance and a representation (a reference pattern or a reference model) of the vocabulary words obtained by a training phase. Two main distortions are generally observed between them. The first one is a non-linear warping of the time scale, which can be coped with by using the dynamic time warping (DTW) technique. The second difference concerns the pronunciation itself in that a dissimilarity remains in spite of optimal restoration of the time scale.

If the reference patterns or both the input utterance and the reference patterns are represented as sequences of VQ code sequences instead of spectral parameter sequences, the amount of computation and storage for the reference patterns can be greatly reduced [1][2]. Here, the VQ is also implicitly used as a clustering technique to generate good, efficient and reliable templates.

In the SPLIT (strings of phoneme-like templates) system, the first word recognition system based on this technique [1], phoneme-like templates (codewords, prototype vectors) are generated by clustering a set of spectral parameters in training utterances, and each word is represented as a sequence of templates. Since the spectral distance calculation is performed between input utterance frames and phoneme-like templates, the amount of calculation does not depend on the vocabulary size. This method is, therefore, especially effective for speaker-dependent large vocabulary word recognition, as well as speaker-independent word recognition using multiple templates.

2.2 Discrete HMM

Recently, the hidden Markov model (HMM) [3]-[9] has become more popular than the DTW technique. The HMM has the capability of modeling both of the two main distortions described above: the time warping and the pronunciation variations. The HMM which uses VQ-based discrete spectral density is called "discrete HMM" in contrast to "continuous HMM" which uses continuous spectral density functions.

A typical structure of a discrete HMM is based upon a left-to-right Markov chain. The observed spectral sequence of an utterance is assumed to be a stochastic function of the state sequence of the Markov chain. The state sequence itself is unobservable (hidden). The parameters characterizing the HMM are the number of states, the state transition matrix, and the observation probability functions.

Training of the HMM, that is, choosing the parameters to optimally match the observed spectral sequences can be performed with the Baum-Welch algorithm. In the recognition phase for an unknown input, the probability that the observed spectral sequence is generated from an HMM for each vocabulary word is calculated with the forward-backward or the Viterbi algorithm which is similar to the DTW algorithm. The word with the highest probability is selected as the correct recognition.

2.3 Semi-Continuous HMM (SCHMM) and Fuzzy-VQ-Based HMM (FVQHMM)

Although the discrete HMM has various advantages, such that it can model events with any distribution provided enough training data exist, it has a serious problem of causing quantization errors. That is, the VQ operation partitions the spectral space into separate regions according to some distortion measure. Solving this problem by increasing the codebook size creates another problem in that huge amounts of training utterances are necessary and that the amount of computation in the recognition stage becomes very large.

From this point of view, the continuous HMM has been introduced, and is becoming more popular than the discrete HMM. However, the continuous HMM also has a problem in that mixtures of a large number of probability density functions considerably increase not only the computational complexity, but also the number of free parameters that need to be reliably estimated. To cope with

1058-6393/91 $1.00 © 1991 IEEE

these problems, the semi-continuous HMM (SCHMM) [10] and the fuzzy-VQ based HMM (FVQHMM) [11][12], which are somewhere in between the discrete and continuous HMMs, have been proposed.

In the SCHMM, the VQ codebook consists of a mixture of continuous probability density functions (for example, each codeword is represented by a mean vector and a covariance matrix), such that the distributions are overlapped, rather than partitioned. The SCHMM has the modeling ability of large-mixture probability density functions. In addition, the number of free parameters and the computational complexity can be reduced in comparison with the continuous HMM, since all of the probability density functions are tied together in the codebook. The SCHMM thus provides a solution to the conflict between detailed spectral modeling and insufficient training data. The VQ codebook can also be optimized together with the HMM parameters in terms of the maximum likelihood criterion.

The fuzzy VQ is based on a fuzzy K-means method used in pattern recognition. Unlike the standard VQ that generates the index of a single codeword that best matches an input vector, the fuzzy VQ makes a soft decision about which codeword is closest to the input vector. It generates an output vector whose components indicate the relative closeness of each codeword to the input.

It is important that the model, trained by limited data, is robust against variations, such as noise, recording conditions, and speaking styles. Robustness of six types of phoneme-HMMs against speaking-style variations was examined [13]. The six types were discrete HMM, FVQHMM, and single-Gaussian and mixture-Gaussian HMMs with either diagonal or full covariance matrices. Eighteen Japanese-consonant recognition experiments were performed using isolated word utterances, phrase-by-phrase utterances, and sentence utterances. The FVQHMM, the mixture-Gaussian HMM with diagonal covariance matrices and the single Gaussian HMM with full covariance matrices displayed better results than the other three types, when different speaking-style utterances were used in training and testing.

2.4 VQ-Distortion-Based Recognition

In all of the above-mentioned methods, the vector quantizers have been incorporated with a DTW- or HMM-based time alignment mechanism. Recently, simpler recognition algorithms based more directly on VQ have been proposed, in which each word in the vocabulary has its own codebook, designed by clustering the training data for that word. The input test sequence is then encoded by the codebook for each word and the word corresponding to the minimum distortion codebook is chosen as the recognizer output [14][15]. Although this method produced good results with a small amount of computation, especially as a method of preprocessor in large vocabulary recognition [16], the lack of temporal information caused problems in recognizing similar words.

Several variations of the VQ-based recognizer have therefore been proposed to include temporal information. One approach is to divide the words into several sections (or states) and design a sequence of codebooks, one for each section, for each word (multi-section VQ) [17].

Another approach is a conditional histogram approach, which incorporates the relative likelihoods that certain codewords follow others into the distortion measure [18]. The distortion between the input frame and a codeword is defined as the weighted sum of the ordinary spectral distortion and the negative logarithm of the conditional probability of getting this codeword given the predecessor in the same codebook.

Another technique of incorporating temporal information into the codebook has been proposed, in which cepstral and delta-cepstral parameters (short-time regression coefficients) are jointly used as spectral features [19]. Figure 1 shows the block diagram of a word recognizer incorporating a VQ preprocessor and an HMM- or DTW-based postprocessor. A universal codebook for the above-mentioned parameters is constructed based on a multi-speaker, multi-word database, and a separate codebook is designed as a subset of the universal codebook for each word in the vocabulary. These codebooks are used for preprocessing to eliminate word candidates whose distance scores are large. A discrete HMM- or DTW-based recognizer using the universal codebook then resolves the choice among the remaining word candidates. Recently this method was expanded to include hierarchical delta-cepstral parameters extracted over multiple time lengths [20].

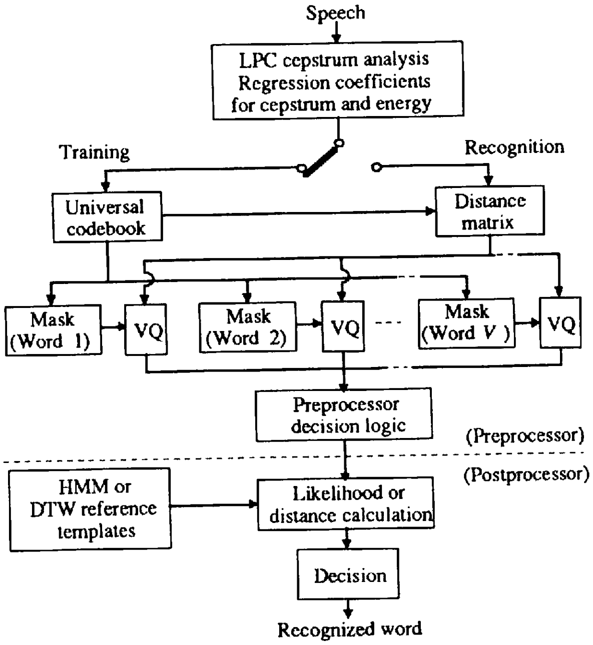

Fig. 1 - Block diagram of a word recognizer incorporating a VQ preprocessor and an HMM-/DTW-based postprocessor.

3. Learning VQ (LVQ)

Learning vector quantization (LVQ) [21][22] is a classifier that is closely linked to Kohonen's work on self-organizing feature maps [23]; the main difference is that LVQ is concerned with finding good category boundaries, while the self-organizing feature maps are designed to find reference vectors that are centroids of the input vectors. The learning in LVQ is, therefore, supervised, while that in self-organizing feature maps is unsupervised.

In LVQ, each category to be learned is assigned a number of reference vectors. Initial configurations of the reference vectors are usually obtained by using the traditional K-means clustering procedure. LVQ training then tries to adjust these positions so that each input vector has a reference vector of the right category as its closest reference vector. Kohonen proposed two versions of LVQ: LVQ1 and LVQ2. The difference between them is the way they select the reference vectors to be adapted.

The adaptation rule for LVQ1 is as follows. If the reference vector closest to the input vector belongs to the same category as the input vector, it is moved closer to the input vector, in proportion to the distance between the two vectors. If the closest reference vector belongs to a category other than that of the input vector, it is moved away, again in proportion to the distance between the two vectors. LVQ2 requires that a number of conditions be met before vector adaptation can occur. These conditions allow the system to pay closer attention to the decision lines of a given categorization problem.

4. Speaker Adaptation by VQ-Codebook Mapping

A number of approaches have been tried in an effort to build speaker-independent recognition systems, typically under HMM-based frameworks. However, since the distributions of feature parameters across speakers are very broad, it is difficult to separate phonemes using speaker-independent methods. Speaker adaptation is a method of automatically adapting reference templates to each new speaker or normalizing interspeaker variations in input utterances.

Since a VQ codebook represents the distribution of given samples in a multi-dimensional spectral space in a non-parametric way, the relationships between spectral distributions of a reference speaker and a new speaker can be represented by correspondences between codewords associated with them. Using mapping rules based on the correspondences, spectra of a new speaker can be adapted to the reference speaker or vice versa. Individual variations on how a word is uttered are modeled by an HMM or multiple sequences of codebook entries in the word dictionary. Speaker-adaptation methods are generally classified into supervised (text-dependent) methods in which training words or sentences are known, and unsupervised (text-independent) methods in which arbitrary utterances can be used.

4.1 Supervised Learning

For supervised adaptation [24], the mapping rules are obtained through DTW or the Viterbi algorithm. First, utterances of a reference speaker are used to create an initial codebook. These utterances are then vector-quantized, that is, converted into sequences of codewords. In the training stage, training utterances of a new speaker are converted into code sequences and time-aligned with the same word or sentence uttered by the reference speaker. The spectral mapping function between the codewords of these two speakers is obtained from alignment functions, that is, the correspondences between the time axes.

Each codeword is included in various words, and each codeword of the reference speaker corresponds to various codewords of the new speaker. Thus, a histogram of correspondences between codewords of the reference speaker and the new speaker, that is, a histogram of co-occurrences of codewords, is calculated using the alignment results of all training words or sentences. The mapping function is weighted by the histogram to find the best correspondence rule. In the recognition stage, input speech is vector-quantized and mapped to the reference speaker's spectrum using the mapping rules at every frame. The similarity between the normalized input speech and each word of the reference speaker is then calculated and used in the recognition decision.

In HMM-based recognition, probabilistic spectral mapping from the reference speaker's spectral space to that of the new speaker has also been investigated [25]. The transformation matrix, which represents the conditional probability of a codeword of the new speaker, given a codeword of the reference speaker, is computed by applying a modified forward-backward algorithm to training utterances.

An approach to speaker adaptation for a large-vocabulary HMM-based speech recognition system has also been tried [26]. The approach is based on the use of a stochastic model, called the "speaker Markov model". The model indicates which codeword of the new speaker is likely to occur and what spectral parameters are generated by the reference speaker if, at the same time, a certain codeword is generated by the new speaker.

4.2 Unsupervised Learning

Figure 2 is a block diagram of an unsupervised codebook adaptation method [27]. The idea of this method is based on an adaptation algorithm for a segment vocoder [28]. First, an initial codebook and a VQ-indexed word dictionary are prepared. The initial codebook is produced by clustering the voices of multiple speakers, and commonly serves as the initial condition for each new speaker.

In the adaptation process, a set of spectra from the

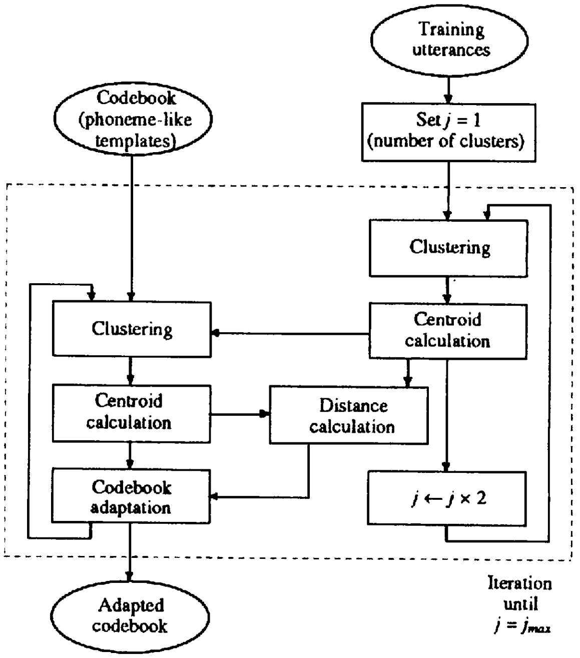

Fig. 2 - Block diagram of an unsupervised codebook adaptation method.

training utterances of a new speaker and the reference codebook elements are clustered hierarchically in an increasing number of clusters. Using the deviation vectors between centroids of the training spectra clusters and the corresponding codebook clusters, either codebook elements or input frame spectra are shifted so that the corresponding centroids coincide. Continuity between adjacent clusters is maintained by determining the shifting vectors to be the weighted-sum of the deviation vectors of adjacent clusters. Adaptation is thus performed hierarchically from global to local individuality. Several modifications to the adaptation method have also been investigated [29].

5. VQ-Distortion-Based Speaker Recognition

VQ-based speaker recognition methods, which are similar to the VQ-distortion-based speech recognition methods, have been investigated. Speaker-specific codebooks are produced in the training stage by clustering the spectral distribution of each reference speaker. In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker, and the mean value or the distribution of the quantization error over the entire speech interval is calculated. These values obtained using each reference codebook are examined to make the recognition decision.

Speaker recognition methods can be classified into text-dependent and text-independent methods. The former require the speaker to issue a predetermined utterance, whereas the latter do not rely on a specific text being spoken.

5.1 Text-Dependent Recognition

In the text-dependent speaker recognition, three VQ approaches have been investigated: single section VQ (normal VQ), multi-section VQ, and matrix quantization (MQ) [30]. The multi-section VQ and the MQ approaches are two different ways of incorporating temporal information into the recognition process. Multi-section VQ models a source by dividing it into several independent, time-ordered subsources. On the other hand, MQ models an utterance with a single codebook that contains an unordered set of time-ordered speech spectrum sequences. Although speaker verification performances of the three approaches when using only a single digit per speaker were similar, the multi-section VQ approach did best when 10- or 5-digit sets were used.

5.2 Text-Independent Recognition

A method using two VQ codebooks, containing the cepstral and delta-cepstral parameters, has been examined [31]. The VQ codebooks for each speaker are constructed using the isolated utterances of 10 digits. An unknown speaker then says any one of the 10 digits. Therefore, this method is text-independent but vocabulary-dependent. Distances (distortions) from test vectors to the two VQ codebooks are optimally combined and averaged over the test utterance to make a final recognition decision. The experimental results show that since the cepstral and delta-cepstral parameters are relatively uncorrelated, they provide complementary information for speaker recognition. They also show that the transitional representations and performance are relatively resistant to simple transmission channel variations.

A method using a single codebook for long feature vectors consisting of instantaneous and transitional features representing both cepstral and pitch characteristics has recently been investigated [32]. Figure 3 is a block diagram of the recognition system. Three key techniques were introduced to cope with any temporal and text-dependent spectral variations. First, either an ergodic HMM or a voiced/unvoiced decision was used to classify input speech into broad phonetic classes. Second, a new distance measure, Distortion-Intersection Measure (DIM), was introduced for calculating VQ distortion of input speech using speaker-specific codebooks. DIM is characterized by selective matching using only a stable subset of test speech in the distortion calculation. Third, a new normalization. method, Talker Variability Normalization (TVN), was introduced. TVN normalizes parameter variation taking both inter- and intra-speaker variability into consideration. TVN emphasizes feature parameters that have relatively large inter-speaker variability and small intra-speaker variability. The combination of the three techniques provides highly accurate speaker identification.

A connectionist approach based on the LVQ algorithm has also recently been tried [33].

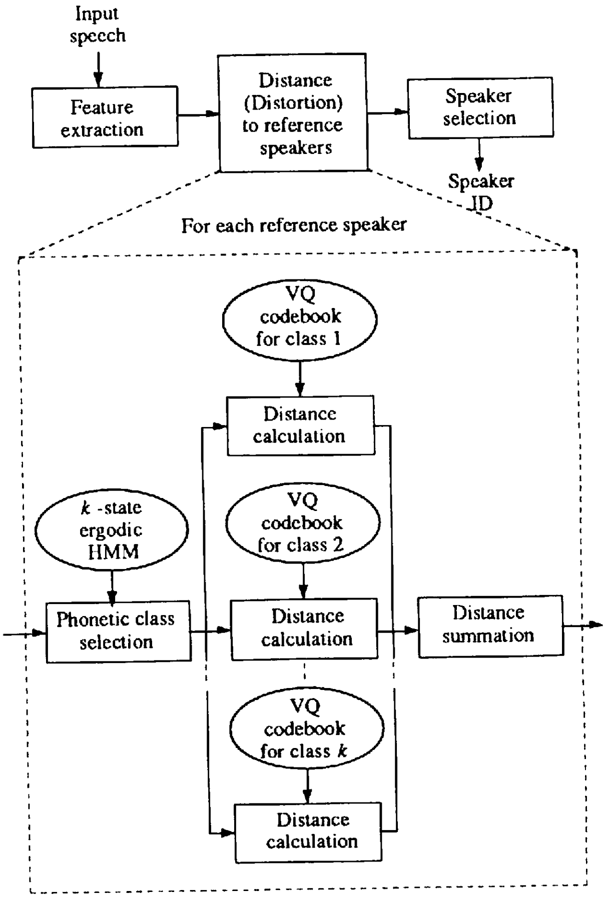

Fig. 3 - Block diagram of text-independent speaker recognition incorporating broad phonetic classification.

6. Summary

This paper reviewed various methods of applying the

VQ technique to speech/speaker recognition. Not only has it reduced the amount of computation and storage, the VQ technique has also created new ideas of solving various problems in speech/speaker recognition, such as speaker adaptation using codebook mapping. Several techniques, such as fuzzy VQ and semi-continuous modeling of HMM, have also been investigated to cope with the quantization distortion problem. The VQ technique has the potential to produce new ideas in speech/speaker recognition in combination with new techniques, such as neural networks and statistical approaches.

References

[1] N. Sugamura, K. Shikano and S. Furui (1983), "Isolated word recognition using phoneme-like templates," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Boston, 16.3.

[2] H. Bourlard, C. J. Wellekens and H. Ney (1984), "Connected digit recognition using vector quantization," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, San Diego, 26.10.

[3] L. R. Rabiner, S. E. Levinson and M. M. Sondhi (1982), "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," Bell System Tech. J., 62, 4, pp.1075-1105.

[4] R. Billi (1982), "Vector quantization and Markov source models applied to speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Paris, France, pp. 574-577.

[5] M. A. Bush and G. E. Kopec (1985), "Network-based connected digit recognition using vector quantization," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tampa, 31.1.

[6] H. Bourlard, Y. Kamp and C. J. Wellekens (1985), "Speaker dependent connected speech recognition via phonemic Markov models," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tampa, 31.5.

[7] Y. L. Chow, et al. (1987), "BYBLOS: The BBN continuous speech recognition system," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas, 3.7.

[8] K.-F. Lee (1989), "Automatic speech recognition - The development of the SPHINX system," Kluwer Academic Publishers.

[9] L. R. Bahl, et al., (1989), "Large vocabulary natural language continuous speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, Scotland, S9.8.

[10] X. D. Huang, Y. Ariki and M. A. Jack (1990), "Hidden Markov models for speech recognition," Edinburgh Univ. Press.

[11] H.-P. Tseng, M. J. Sabin and E. A. Lee (1987), "Fuzzy vector quantazation applied to hidden Markov modeling," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas, 15.5.

[12] T. Hanazawa, K. Kita, S. Nakamura, T. Kawabata and K. Shikano (1990), "ATR HMM-LR continuous speech recognition system," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S2.4.

[13] T. Matsuoka and K. Shikano (1991), "Robust HMM phoneme modeling for different speaking styles," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto, Canada, S5.4.

[14] A. Buzo, H. Martinez and C. Rivera (1982), "Discrete utterance recognition based upon source coding techniques," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Paris, France, pp. 539-542.

[15] J. E. Shore and D. K. Burton (1983), "Discrete utterance speech recognition without time alignment," IEEE Trans. Inform. Theory, 29, 4, pp. 473-491.

[16] K.-C. Pan, F. K. Soong and L. R. Rabiner (1985), "A vector-quantization-based preprocessor for speaker-independent isolated word recognition," IEEE Trans. Acoust., Speech, Signal Processing, 33, 3, pp. 546-560.

[17] D. K. Burton, J. E. Shore and J. T. Buck (1985), "Isolated-word speech recognition using multisection vector quantization codebooks," IEEE Trans. Acoust., Speech, Signal Processing, 33, 4, pp. 837-849.

[18] S.-S. Huang and R. M. Gray (1988), "Spellmode recognition based on vector quantization," Speech Communication, 7, 1, pp. 41-53.

[19] S. Furui (1987), "A VQ-based preprocessor using cepstral dynamic features for large vocabulary word recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas, 27.2.

[20] S. Furui (1990), "On the use of hierarchical spectral dynamics in speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S15a.10.

[21] T. Kohonen (1988), "Self-organization and associative memory (2nd ed.)", Springer Verlag, pp.199-202.

[22] E. McDermott and S. Katagiri (1991), "LVQ-based shift-tolerant phoneme recognition," IEEE Trans. Signal Processing, 39, 6, pp. 1398-1411.

[23] T. Kohonen, G. Barna and R. Chrisley (1988), "Statistical pattern recognition with neural networks: Benchmarking studies," Proc. IEEE Int. Conf. Neural Networks, I, pp.61-68.

[24] K. Shikano, K.-F. Lee and R. Reddy (1986) " Speaker adaptation through vector quantization," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, Japan, 49.5.

[25] R. Schwartz, Y.-L.Chow and F. Kubara (1987), "Rapid speaker adaptation using a probabilistic spectral mapping," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas, 15.3.

[26] G. Rigoll (1989), "Speaker adaptation for large vocabulary speech recognition systems using 'speaker Markov models'," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, Scotland, S1.2.

[27] S. Furui (1989), "Unsupervised speaker adaptation based on hierarchical spectral clustering," IEEE Trans. Acoust., Speech, Signal Processing, 37, 12, pp. 1923-1930.

[28] Y. Shiraki and M. Honda (1990), "Speaker adaptation algorithms based on piece-wise moving adaptive segment quantization method," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S12.5.

[29] S. Furui (1989), "Unsupervised speaker adaptation method based on hierarchical spectral clustering," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, Scotland, S6.9.

[30] D. K. Burton (1987), "Text-dependent speaker verification using vector quantization source coding," IEEE Trans. Acoust., Speech, Signal Processing, 35, 2, pp.133-143.

[31] F. K. Soong, and A. E. Rosenberg (1988), "On the use of instantaneous and transitional spectral information in speaker recognition," Trans. Acoust., Speech, Signal Processing, 36, 6, pp.871-879.

[32] T. Matsui and S. Furui (1991), "A text-independent speaker recognition method robust against utterance variations," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto, Canada, S6.3.

[33] Y. Bennani, F. Fogelman Soulie and P. Gallinari (1990), "A connectionist approach for automatic speaker identification", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S5.2.