SPEAKER CHANGE DETECTION AND SPEAKER CLUSTERING USING VQ DISTORTION FOR BROADCAST NEWS SPEECH RECOGNITION

Kazumasa MORI and Seiichi NAKAGAWA

Dept. of Information and Computer Sciences,
Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, Aichi-ken, 441-8580 JAPAN

ABSTRACT

This paper addresses the problem of the detection of speaker changes and clustering speakers when no information is available regarding speaker classes or even the total number of classes. We assume that no previous information on speakers is available (no speaker model, no training phase) and that people do not speak simultaneously. The aim is to apply speaker grouping information to speaker adaptation for speech recognition.

We use Vector Quantization (VQ) distortion as the criterion. A speaker model is created from successive utterances as a codebook by a VQ algorithm, and the VQ distortion is calculated between the model and an utterance.

The result was given by the experiment on speaker detection and speaker clustering. The speaker change detection experiment was compared with results by Generalized Likelihood Ratio (GLR) and Bayesian Information Criterion (BIC). We show the superiority of our proposed method.

1. INTRODUCTION

This paper addresses the problem of the detection of speaker changes and clustering speakers when no information is available regarding speaker classes or even the total number of classes. We assume that no previous information on speakers is available (no speaker model, no training phase) and that people do not speak simultaneously. The aim is to apply speaker grouping information to speaker adaptation for speech recognition.

In a general task, some speakers' utterances are included in speech documents. Applying the same model to different speaker's utterances is detrimental to recognition performance. Thus, speaker detection and adaptation are effective before recognition. Previous studies have dealt with the problem of speaker detection. For example, several distance measures have been used to calculate speaker differences, such as GLR [1] [2] and BIC [3] [2] [4]. These BIC studies mostly experimented with a full covariance matrix. In this paper, we describe the expansion of BIC for multiple mixtures (GMM) and the evaluation.

When the acoustic characteristics are unknown, unsupervised adaptation techniques can be effective in improving performance. Such methods are more effective as the amount of adaptation data increases, so it is of interest to cluster segments from the same speaker and condition [5] [6]. The goal of data partitioning is to divide the acoustic signal into homogenous segments, and to associate appropriate labels with the segments.

In the conventional approach, speaker models are represented by Gaussian Mixture Models (GMM). However, in order to make a reliable GMM, much speech data for more than 10 seconds is required. A Vector Quantization (VQ) distortion technique is a special case of a GMM, therefore it is robust and preferable to a GMM for a short utterance. Thus, we use VQ distortion as the criterion. A speaker model is created from successive utterances as a codebook by a VQ algorithm, and the VQ distortion is calculated between the model and an utterance.

2. SYSTEM OVERVIEW

The block diagram in Figure 1 shows the major components of the broadcast news speech recognition system developed in our laboratory.

The system has used speaker-adapted acoustic models for announcers and speaker-independent acoustic models for other speakers such as reporters. The techniques are as follows: (1) Acoustic models for announcers are adapted in training mode in advance; (2) the speaker for an utterance is identified in test mode; and (3) if the utterance is identified as voices uttered by one of the announcers, the acoustic model corresponding to the identified speaker is used for speech recognition. On the other hand, for the case of rejected speakers (i.e., not announcers, but for those we call "others"), the system uses the adapted model corresponding to the result estimated by the speaker change detection and the speaker clustering, and this utterance is used for the adaptation.

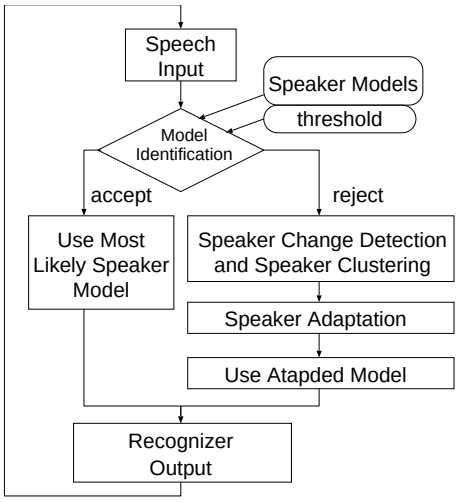

Fig. 1. Block diagram of system.

3. CRITERION FOR SPEAKER CHANGE DETECTION

For the two portions of parameterized signals (sequences of acoustic vectors) A and B shown in Figure 2 , in this section we describe how to determine whether they are identical or not.

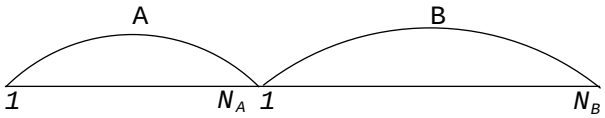

Fig. 2. Acoustic segments.

3.1. GAUSSIAN SPEAKER MODELS

A GMM is weighted sum of M component densities and is given by

$$p(x) = \sum_{i=1}^{M} c_i \frac{1}{2\pi^{d/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\}, \quad (1)$$

where x is a d-dimensional random vector, $c_i, i = 1, ..., M$ is the mixture weight, μ_i the mean vector and Σ_i the covariance matrix. The mixture weights satisfy the constraint

$$\sum_{i=1}^{M} c_i = 1. \quad (2)$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$c_i, \mu_i, \Sigma_i, \quad i = 1, ..., M. \quad (3)$$

Eq.(1) is approximated to

$$p(x) = \max_i c_i \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\}, \quad (4)$$

which equals

$$-\log p(x) = \min_i \left\{-logc_i + \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma_i| + \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\}. \quad (5)$$

In our system, GMMs are used for speaker identification [7].

3.2. Generalized Likelihood Ratio

The Generalized Likelihood Ratio (GLR) is defined by

$$R = \frac{L(A\cup B; N(\mu_{A\cup B}, \Sigma_{A\cup B}))}{L(A; N(\mu_A, \Sigma_A)) \cdot L(B; N(\mu_B, \Sigma_B))}, \quad (6)$$

where $L(X; N(\mu_X, \Sigma_X))$ is the likelihood of X for the Gaussian model $N(\mu_X, \Sigma_X)$. A high value of R signifies that the one multi-Gaussian modeling best fits the data. By contrast, a low value of R indicates that a speaker change is detected between A and B.

3.3. Bayesian Information Criterion

The BIC is a likelihood criterion penalized by the model complexity. The BIC value is determined by

$$BIC = \log L(A; N(\mu_A, \Sigma_A)) - \frac{1}{2}\lambda\left(d + \frac{d(d+1)}{2}\right)\log N_A. \quad (7)$$

The variation of the BIC value between the two models is then given by

$$\Delta BIC = -\frac{N_A+N_B}{2}\log|\Sigma_{A\cup B}| + \frac{N_A}{2}\log|\Sigma_A| + \frac{N_B}{2}\log|\Sigma_B| + \frac{1}{2}\lambda\left(d + \frac{d(d+1)}{2}\right)\log(N_A+N_B), \quad (8)$$

where d is the dimension of the acoustic space, and λ is the penalty factor. A high value of ΔBIC indicates that the one multi-Gaussian models best fit the data.

We extend the above BIC to GMM as follows:

$$\Delta BIC = -\sum_{m=1}^{M}\frac{N_{(A\cup B)_m}}{2}\log|\Sigma_{(A\cup B)_m}| + \sum_{m=1}^{M}\frac{N_{A_m}}{2}\log|\Sigma_{A_m}| + \sum_{m=1}^{M}\frac{N_{B_m}}{2}\log|\Sigma_{B_m}| + M\cdot\frac{1}{2}\lambda\left(d + \frac{d(d+1)}{2}\right)\log(N_A+N_B), \quad (9)$$

where M is the number of mixtures, and N_{X_m} means the number of samples assigned as the m-th mixture distribution. We used Eq.(4) instead of Eq.(1), and neglected the mixture weight.

3.4. GMM Likelihood Measure

The GMM Likelihood Measure (GMM-L) is defined by

$$GMM{-}L \quad = \quad \frac{1}{N_B} \log L(B; N(\mu_A, \Sigma_A)), \quad (10)$$

where we used Eq.(4) instead of Eq.(1). A low value of $GMM{-}L$ indicates that a speaker change occurs between A and B.

The GMM Likelihood Measure is simpler than the BIC and GLR, and the asymmetry between utterances differs from many other criteria for the speaker change detection.

3.5. VQ Distortion Measure

The VQ Distortion Measure is defined by

$$D = \frac{1}{N_B} \sum_{j=1}^{N_B} \min_{1 \le i \le |C_A|} d(C_A(i), B(j)), \quad (11)$$

where C_X is the VQ codebook created by acoustic vector sequence X, $|C_X|$ is the codebook size of C_X, and $D(C_X(i), Y(j))$ is the Euclidean distance between the i-th code vector of the codebook C_X and j-th frame of the vector Y.

If we substitute $1/M$ for c_i and the identity matrix I for Σ, respectively, the GMM likelihood measure and the VQ distortion measure become identical. So we can regard the VQ distortion measure as a special case of GMM likelihood measure [8].

This measure is robust and reliable for a short utterance because the number of estimated parameters is reduced remarkably.

4. ALGORITHM

4.1. SPEAKER CHANGE DETECTION WITH VQ DISTORTION MEASURE

Given the utterances $X(1), ..., X(i), ..., X(N)$, the speaker change detection is followed that

0. Initially, i = 1
1. Calculate $D(C(i), X(i+1))$.
2. If $D(C(i), X(i+1)) > threshold$, we assume that the speaker change occurred at the boundary of $X(i)$ and $X(i+1)$.
3. $i = i + 1$, and while $i < N$ return to (1).

If the successive utterances are found to be utterances spoken by the same speaker, we use a set of codebooks, each of which corresponds to the respective utterance. The distortion between the set and the next utterance is defined as the average distortion between each codebook and the utterance. We call this multi-codebooks.

4.2. SPEAKER CLUSTERING WITH VQ DISTORTION MEASURE

The speaker clustering is performed as follows:

0. Create a codebook by $X(1)$. $i = 1$.
1. VQ distortion is calculated between $X(i+1)$ and a codebook of each cluster.
2. If the "nearest cluster" within the threshold is found, the utterance is merged with it; otherwise a new cluster is created. Only the $X(i+1)$ utterance belongs to the new cluster.
3. $i = i + 1$, and while $i < N$ return to (1).

5. EXPERIMENTS AND RESULTS

5.1. DATABASE

We used a subset of clean speech (175 utterances in total) from the NHK (Japan Broadcasting Corporation) broadcast news speech database for the evaluation experiments. In the database, people do not speak simultaneously. There are 27 turns for speaker change. The subset contains 2 announcers (I,J) and 8 other speakers (A,B,C,D,E,F,G,H). The number of announcers in training data is 6 (I,J,K,L,M,N). Therefore our system has 6 speaker GMMs. The utterance length is between 0.5s and 17.9s and the average length is 7.6s.

Table 1 shows the speech analysis conditions. Table 2 shows the result of the speaker identification. There was no confusion between announcers and 17% confusion between announcers and "others."

Table 1. Speech analysis conditions

Sampling frequency	12 kHz
Pre-emphasis	$1 - 0.98z^{-1}$
Hamming window width	21.33 ms (256 points)
Frame period	8 ms (96 points)
LPC analysis order	14-th
Feature parameters	KL-20 dimensional coefficients [9] cepstrum

Table 2. Confusion matrix of speaker identification

speakers\models	I	J	K	L	others	total
I	47	0	0	0	12	59
J	0	42	0	0	2	44
others	7	3	1	5	56	72
total	54	45	1	5	70	175

5.2. SPEAKER CHANGE DETECTION

Recall(RCL) , precision(PRC) and F are defined as

$$RCL = \frac{number\ of\ correctly\ found\ boundaries}{total\ number\ of\ boundaries}, \quad (12)$$

$$PRC = \frac{number\ of\ correctly\ found\ boundaries}{number\ of\ hypothesized\ boundaries}, \quad (13)$$

$$F = \frac{RCL \cdot PRC}{\alpha \cdot RCL + (1-\alpha) \cdot PRC}, \quad (14)$$

where α was set to 0.5 in this evaluation.

Table 3 shows the best result in terms of F-measure. We find that our proposed VQ-based method is superior to BIC, GLR, GMM-L and that the BIC of GMM with 2 mixtures is superior to a single Gaussian-based BIC. When we used multiple utterances uttered by the same speaker (multiple codebooks), the F-measure was improved remarkably. Next, we replaced a low probability or large distance to a constant value (acoustic model's back-off [11]), because of a few training samples. The results were remarkably improved, and the VQ-based method was still the best.

Table 3. Best result of F-measure.

(a) standard local probability / distance

criterion			F-measure
GLR	full cov.	1 mixture	75.9
		2 mixtures	73.5
		4 mixtures	73.7
	diagonal cov.	2 mixtures	74.5
		4 mixtures	72.4
		8 mixtures	74.6
		16 mixtures	75.4
BIC	1 mix. full cov.	$\lambda = 7, 8, 9, 10$	80.0
	2 mix. full cov.	$\lambda = 6, 7$	81.4
	4 mix. full cov.	$\lambda = 6$	75.0
GMM-L	full cov.	2 mixtures	78.8
		4 mixtures	75.0
	diagonal cov.	4 mixtures	78.6
		8 mixtures	80.0
		16 mixtures	77.2
VQ	codebook size =32		83.1
	codebook size=64		83.8
	codebook size=128		83.9
VQ	multi-codebooks,codebook size =64		89.3

(b) back-off local probability / distance

criterion			F-measure
GLR	full cov.	4 mixture	75.5
	diagonal cov.	16 mixture	77.4
GMM-L	full cov.	4 mixture	87.1
	diagonal cov.	16 mixture	88.1
VQ	codebook size = 32		90.0

5.3. SPEAKER CLUSTERING

Rejected utterances by speaker identification, that is, utterances regarded as "others" were clustered into some classes. Fig.4 shows the result of speaker clustering with VQ distortion measure. The cluster C_0, C_1, C_4, C_5, C_6 and C_{10} is purely clustered. The Clustering Efficiency (CE) [10] was 0.56.

6. CONCLUSIONS AND FUTURE WORK

We conclude that using VQ distortion between utterances is more effective than the GLR and BIC for the detection of speaker change and the speaker clustering. Furthermore, the recognition rate would be improved with the adaptation of speaker models by an MLLR and MAP adaptation technique using speaker clustering.

Table 4. Clustering result.

\input / output	others								announcer		total
	A	B	C	D	E	F	G	H	I	J	
C_0	6	0	0	0	0	0	0	0	0	0	6
C_1	0	0	0	0	0	0	0	0	1	0	1
C_2	0	10	0	0	0	0	0	0	5	0	15
C_3	0	5	0	0	0	0	0	0	3	0	8
C_4	0	0	0	0	0	0	0	0	3	0	3
C_5	0	0	3	0	0	0	0	0	0	0	3
C_6	0	0	0	4	0	0	0	0	0	0	4
C_7	0	0	1	0	9	0	0	0	0	0	10
C_8	0	0	1	0	0	2	0	0	0	1	4
C_9	0	0	0	0	0	4	7	1	0	1	13
C_{10}	0	0	0	0	0	0	0	3	0	0	3
total	6	15	5	4	9	6	7	4	12	2	70

ACKNOWLEDGMENT

This research has been partially supported by the NHK Science and Technical Research Laboratories.

REFERENCES

[1] J.-F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, C. Wellekens, "A speaker tracking system based on speaker turn detection for NIST evaluation", Proc. ICASSP 2000, pp.1177-1180 (2000)

[2] P. Delacourt, D. Kryze, C. J. Wellekens, "Detection of speaker changes in an audio document", Proc. EuroSpeech '99, pp.1195–1198 (1999)

[3] S. Chen, P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion", Proc. DARPA Speech Recognition Workshop pp.127-132 (1998)

[4] B. Zhou, J. H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian Information Criterion", Proc. ICSLP 2000, pp 714–717 (2000)

[5] A. Solomonoff, A. Mielke, M. Schmidt, H, Gish, "Clustering speakers by their voices", Proc. ICASSP ' 98, pp.757-760 (1998)

[6] J-L. Gauvain, L. Lamel and G. Adda, "Partitioning and transcription of broadcast news data", Proc. ICSLP '99, pp.1335-1338 (1998)

[7] K. Markov, S. Nakagawa: "Text-independent speaker recognition using non-linear frame likelihood transformation", Speech Communication, Vol.24, pp.193-209 (1998)

[8] S. Nakagawa, H. Suzuki: "A new speech recognition method based on VQ-distortion measure and HMM", Proc. ICASSP '93, pp.676-679 (1993)

[9] S. Nakagawa, K. Yamamoto, "Evaluation of segmental unit input HMM", Proc. ICASSP '96, pp.1439-1442 (1996)

[10] D.A. Reynolds, E. Singer, B.A. Carlson, G.C. O'Leary, J.J. McLaughlin and M.A. Zissman, "Blind clustering of speech utterance based on speaker and language characteristics", Proc. ICSLP '98, pp.3193-3196 (1998)

[11] J. de Veth, B. Cranen, L. Boves, Acoustic Back-off in the Local Distance Computation for Robust Automatic Speech Recognition, Proc. ICSLP, pp.1427-1430 (1998)