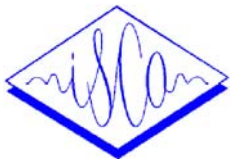

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

N-best GMM's for Speaker Identification

Chakib Tadj [†], *Pierre Dumouchel* [†‡]

[†]École de Technologie Supérieure
1100 rue Notre Dame Ouest
Montreal (Qc) - H3C 1K3 - Canada

[‡]Centre de Recherche Informatique de Montréal
1801, avenue McGill College, bureau 800
Montreal (Qc) - H3A 2N4 - Canada

Yu Fang [◇]

[◇]Institut Universitaire de Technologie
16 Place du commerce
Nun's Island (Qc) - H3E 1H6 - Canada

ABSTRACT

In this paper, we present and compare two alternative post-processing approaches to generate rules decision for text-dependent speaker identification based on Gaussian Mixture Models (GMM). The first approach, a linear programming method, is used to minimize a cost on a combined scores obtained from the N-Best GMM output probabilities. The second, more heuristic, is based on combination of output score probabilities to generate a decision rules. Statistical tools have been developed to explore the relative importance of these approaches on recognition accuracy. Experiments on Spidre database are presented to show the effects of these two approaches on the speaker identification performance (including the number of the N-Best hypothesis and handset variability). The linear programming approach does not show any improvement, however, a combined statistical approaches has demonstrated an improvement of more than 11% comparing to our standard performance system.

Key Words: Speaker Identification , N-Best GMM output probabilities, Heuristic Rules Generation, Linear Programming, Combined Scores.

1. INTRODUCTION

Speaker identification (SID) from the voice's characteristics is increasing with the growing use of speech interaction with computers [3]. SID systems work extremely well with clean speech, but performance degrades considerably with noisy and/or degraded speech as telephone speech [7]. GMM models statically the underlying speech sounds that characterize a person's voice and so is capable of high performance for short test utterances independent of the spoken test. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily shaped densities, it can be viewed as a single state HMM with a Gaussian mixture observation density.

Several works and research have been done during these last years in the context of speaker identification, to improve the performance of the systems [1, 2, 5, 8]. Unfortunately, the decision rules used in these systems are rarely presented. In this paper, we will show the importance of such a post-processing on recognition accuracy. Two alternative post-processing approaches are presented, to generate rules decision for speaker identification, based on Gaussian Mixture Models. Various factors (such as the N-Best output probabilities and handset variability) on the performance of the system are also studied. In section 2, the first approach based on a linear programming method, is used to minimize a cost on a combined scores obtained from the GMM. The second one, an heuristic approach, based on combination of output score probabilities will be presented in section 3. Experiments on Spidre database are showed in section 4. Finally, conclusions and some future problems will be presented in section 5.

2. LINEAR PROGRAMMING APPROACH - FORMULATION

Let c_i, $1 \leq i \leq n$ the output score probability from the i^{th} model of speaker c. n is the number of models for each speaker. These scores are obtained from the N-Best GMM output probabilities.

One way to combine these scores is to weigh them by x_i coefficients in the way that the combined score will be maximized, and minimizing the output score probabilities a_{ij} of all the other speakers. This problem can be formulated in the following way:

$$Max: \quad \sum_{i=1}^{n} c_i x_i$$

$$Constraint: \sum_{j=1}^{n} a_{ij} x_j <= b_i, \qquad i=1,\cdots,m-1 \qquad (1)$$

where n is the number of models for each speaker,

m the number of speakers, c_i and a_{ij} constants to be determined. This formulation is well known as a linear programming problem, which can be solved by the Simplex method [4].

3. HEURISTIC RULES GENERATION

Decision rules are generated from the N-Best GMM output probabilities. Several criteria have been investigated: (a) the best successive scores, (b) range probabilities, (c) weights according to range appearance, (d) weight according to the mean probability and finally (e) the best last criterion.

3.1. The Best Successive Scores

A speaker claimed is accepted by the SID if he is recognized k times with the highest probabilities. More k is high, and more the false acceptation are reduced and false rejection are increased. In our studies, k is set to 2. This implies that the system identify a speaker if he is recognized twice (the corresponding speaker has the two highest output probabilities).

3.2. Range Probabilities

If the range probability between the best score probability and the second best one is higher than a predefined threshold, the claimed speaker is accepted. Practically, the experiment was performed according to:

$$RP = \frac{P_{first\_best}}{P_{second\_best}} \tag{2}$$

where P_{first_best} and P_{second_best} are the first and second best output probabilities of the GMM system respectively. The decision rule is made according to the following scheme:

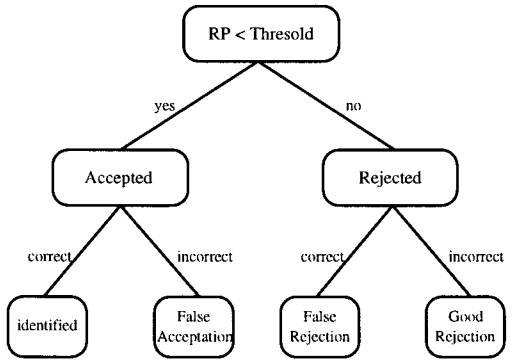

Figure 1: Decision rules scheme in the case of the two best output probabilities

In our experiment, the *Threshold* value is set to 0.996. The output of the two-best GMM probabilities were very close to each other. This make the task more difficult. Of course, more these values are different, and more the performance accuracy is high.

3.3. Weights According to Range Appearance

This criterion weights each score by a weight according to the rank appearance in the list score:

$$w_i = \frac{1}{N_{app_i} N_{can}} \sum_{j=1}^{N_{app_i}} (N_{can} - Rang_{ij}) \tag{3}$$

where N_{app_i} corresponds to the number of appearances of the i^{th} speaker, N_{can} the total number of scores obtained from the N-Best GMM output, $1 \leq N_{can} \leq n * m$, and $Rang_{ij}$ the rank (or order) appearance of i^{th} speaker for the j^{th} model. Note that the best score corresponds to N_{can}=1.

3.4. Weight According to Mean Probability

This criterion evaluates, for each i^{th} speaker, the average probability over all her/his models, according to:

$$w_i = \frac{1}{N_{app_i}} \sum_{j=1}^{N_{app_i}} P_{ij} \tag{4}$$

P_{ij} is the j^{th} output probability for the i^{th} speaker.

3.5. The Best Last

This criterion select the first speaker who has been recognized N_{app_i} times. N_{app_i} is defined in subsection 3.3. In our experiments, N_{app_i} is set to 3, 1, and 2 according to split alternatives used *Split 1*, *Split 2* and *Split 3* respectively (c.f. section 4). Note that, in case *Split 2*, this criterion corresponds to the 1-Best output probability.

3.6. Decision Rules

All the above criteria were combined together, to generates an identification decision. This combination can be summarized as showed in figure 2. Criteria are classified in a decreasing order priority, in the way that if the higher criterion priority is satisfied, a decision is generated, otherwise the next criterion is performed and so on. For example, if the system recognizes twice the same speaker, this speaker is then automatically identified, and there is not need to perform the other criteria. This hypothesis suggest that the system can not make twice the same missclassification.

4. EXPERIMENTS

The experiments use Spidre database. The features are a 26 dimensional vectors consisting of 12 cepstral

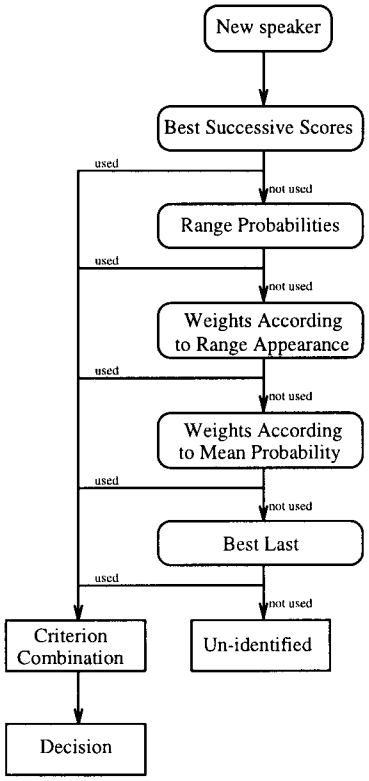

Figure 2: Criteria rules combination

coefficients, 12 Δ coefficients, logarithmic power and Δ logarithmic power. Analysis conditions are listed in table 1. The number of mixture components is set to 34 (16 as static, 16 as dynamic and 2 as energy). Each of the mixture components has a diagonal covariance matrix.

pre-emphasis	1-0.97z^{-1}
window length	25.0 ms
window shift	10.0 ms
MFCC cepstrum order	24
Cepstral coefficient liftering	22
Cepstral mean normalization	yes
Hamming window	yes

Table 1: *Analysis conditions*

The Spidre database consists of 4 conversation halves each from 45 claimant speakers (27 males and 18 females). The 4 conversations from each speaker originate from 3 different handsets (called hereafter *mismatch condition*) with 2 conversations from the same phone number (*match condition*).

In order to not biased the experiments with the match and mismatch conditions, three kind of experiments are performed as shown in table 2. Of course, it is well known that the performance of the systems are decreased in a mismatch condition [7], so the purpose of the experiments here is not to establish this reality. Our goal is to show the behavior of different criteria in different alternatives (match and mismatch).

Split	train	test
Split 1	2 match 1 mismatch	1 mismatch
Split 2	1 match	1 match 1 mismatch
Split 3	1 match 1 mismatch	1 match 1 mismatch

Table 2: *Train and test data description*

In order to study performance of these criteria, the results are compared to two other (and independently) criteria: (a) the best output probability and (b) the ratio of the first 2 best output probabilities (as described in figure 1).

	Criterion	N-Best	Accuracy	Error(%)
Split 1	LP	135 10	36/45 (80.0%) -/45	20.0 -
	BOP	135 10	36/45 (80.0%) 36/45 (80.0%)	20.0 20.0
	RP	135 10	31/45 (68.8%) 31/45 (68.8%)	31.1 31.1
	CS	135 10	38/45 (84.4%) 41/45 (91.1%)	15.6 8.9

Table 3: *Comparison between different criteria for the alternative Split 1. LP=Linear Programming Approach - BOP=Best Output Probability - RP=Range Probabilities - CS=Combined Scores*

Table 3 shows that, the linear programming approach does not improve performance of the system. From Split 1 configuration, the combined criteria performs better than the 1-Best output GMM, and better than considering the only ratio of the 2-Best output probabilities. An other interesting result is the number of the N-Best output chosen as showed in figure 3. The all-Best output has shown a kind of awkness. This can be expected because one of the proposed criteria uses the *best last* output probabilities. In this case the use of all-Best output will certainly not improve the performance because the lastest probabilities are most of the time very small and does not act significantly on the accuracy of the system according to these criteria. The optimal value was found when N-

Best was set to 10. The advantage of this result is double: first the results are improved about 11.2%, and second, the system is faster by avoiding to compute the all possible hypothesis.

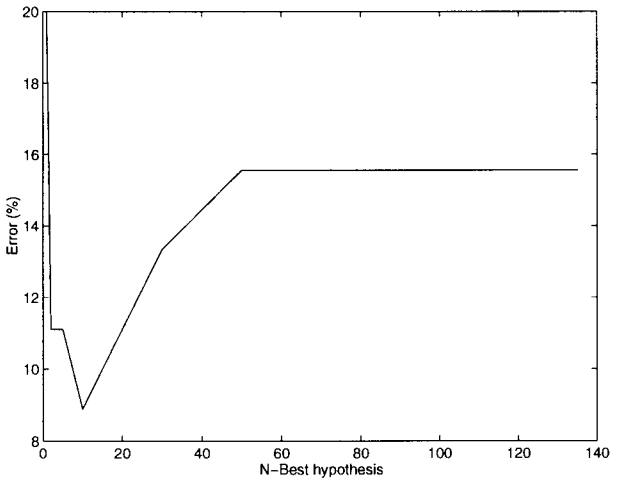

Figure 3: *System accuracy vs. N-Best hypothesis*

In the case of Split 2 and Split 3 sets, the combined criteria does not show, in the most of the time any improvement, as described in table 4, when the number of the N-Best changes. This can be explain by the fact that, in these cases, each speaker has only 1 and 2 models respectively, and therefore the proposed criterion, namely those proposed in sections 3.1, 3.3, 3.4, and 3.5, could not be performed correctly.

Split	Criterion	N-Best	Accuracy		Error
			Match	Mismatch	(%)
Split 2	BOP	45	37/45	21/45	35.5
		10	37/45	21/45	35.5
	RP	45	33/45	15/45	46.6
		10	33/45	15/45	46.6
	CS	45	37/45	27/45	28.9
		10	37/45	27/45	28.9
Split 3	BOP	90	39/45	28/45	25.5
		10	39/45	28/45	25.5
	RP	90	34/45	21/45	38.9
		10	34/45	21/45	38.9
	CS	90	41/45	31/45	20.0
		10	41/45	32/45	18.9

Table 4: *Comparison between different criteria for the 2 alternatives (Split 2, Split 3). BOP=Best Output Probability - RP=Range Probabilities - CS=Combined Scores*

5. CONCLUSIONS AND FUTURE PROBLEMS

In this paper, we have proposed several criteria to improve the speaker identification system's performance. The linear programming approach does not show any improvement, however, a combined statistical approaches has demonstrated an improvement of more than 11% comparing to our standard performance system. To achieve a better accuracy performance, we expect to focus in the future research on both environment problem and the number of mixtures vs. amount of training data, specially when we do not have enough data to train the system.

6. REFERENCES

[1] C. M. Alamo, F. J. Gil, C. T. Munilla and L. H. Gomez, "Discriminative Training of GMM for Speaker Identification", *ICASSP*, pp. 89-92, 1996.

[2] A. Anastasakos, F. Kubala, J. Makhoul and R. Schwartz, "Adaptation to new Microphone Using Tied-Mixture Normalization", *ICASSP*, pp. I-433-436, 1994.

[3] S. Furui, "An Overview of Speaker Technology", *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 1-9, 1994.

[4] D. G. Luenberger, "Linear and Nonlinear Programming", Addison Wesley, 2nd edition, 1984.

[5] R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust Speaker Recognition - A Feature Based Approach", *IEEE Signal Processing*, pp. 58-71, 1996.

[6] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models", *Speech Communication*, pp. 91-108, 1991.

[7] D. A. Reynolds, "The Effect of Handset Variability on Speaker Recognition Performance: Experiments on the Switchboard Corpus", *ICASSP*, pp. 113-116, 1996.

[8] G. Yu and H. Gish, "Identification of Speakers Engaged in Dialog", *ICASSP*, Vol. II, pp. 383-386, 1993.