IDIAP RESEARCH REPORT

Posteriori Probabilities and Likelihoods Combination for Speech and Speaker Recognition

Mohamed Faouzi BenZeghiba [a,b]
Hervé Bourlard [a,b]

IDIAP–RR 04-23

26th April 2004

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), Martigny
[b] Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland

IDIAP Research Report 04-23

Posteriori Probabilities and Likelihoods Combination for Speech and Speaker Recognition

Mohamed Faouzi BenZeghiba Hervé Bourlard

26th April 2004

Abstract. This paper investigates a new approach to perform simultaneous speech and speaker recognition. The likelihood estimated by a speaker identification system is combined with the posterior probability estimated by the speech recognizer. So, the joint posterior probability of the pronounced word and the speaker identity is maximized. A comparison study with other standard techniques is carried out in three different applications, (1) closed set speech and speaker identification, (2) open set speech and speaker identification and (3) speaker quantization in speaker-independent speech recognition.

1 Introduction

Speech signal conveys (among other things) two major types of information, the speech content (text) and the speaker characteristics. Speech recognition systems aim to extract the lexical information from the speech signal. Speaker recognition systems aim to recognize (identify/verify) the speaker. Joint speech and speaker recognition systems aim to recognize (simultaneously) who is speaking and what was said. Such systems have several applications, such as:

1. Speaker identification can be used as a front-end processor to a speech system [1] and vice-versa [2].
2. Performing continuous speaker recognition and knowledge/content recognition [3] [4] [5].
3. Automatic recognition of co-channel speech[2], where more than one speaker is speaking at the same time.

In this paper, a probabilistic approach for the joint speech and speaker recognition is proposed. This approach is based on the combination of a likelihood based speaker-identification with a posteriori probability based speech recognizer. The evaluation of this approach is examined in three applications. Closed set speech and speaker identification (i.e., the access is restricted to speakers enrolled in the system), open set speaker identification (i.e., any speaker can access the system, those that are not enrolled should be rejected) and speaker-quantization for speaker-independent speech recognition[1] (i.e., the speech recognizer associated with the most likely speaker determined by the speaker identification system is used to recognize the utterance pronounced by the speaker). In closed set and open set experiments, our goal is to recognize correctly, both the speaker identity and the command associated with a specific service. A typical application could be the voice dialing system. The combination will be done for every enrolled speaker, making the computational requirements very costly. We will show how to reduce this cost without affecting the recognition performance. Also, we compare the proposed approach in all those situations, with two other standard approaches.

2 Formulation

Our goal is to find the word (command) $\widehat{W}$ from a finite set of possible words $\{W\}$ and the speaker $\widehat{S}$ from a finite set of registered speakers $\{S\}$ that maximize the joint posterior probability $P(\widehat{W}, \widehat{S}|X)$. Formally, this is expressed as follows:

$$\begin{aligned}(\widehat{W}, \widehat{S}) &= \arg\max_{\{W,S\}} P(W, S|X) \\ &= \arg\max_{\{W,S\}} [P(W|S, X) \cdot P(S|X)] \end{aligned} \quad (1)$$

Taking the logarithm, and using Bayes rule with the assumption that the prior probability of the speaker $P(S)$ is uniform over all speakers, equation (1) can be rewritten as:

$$(\widehat{W}, \widehat{S}) = \arg\max_{\{W,S\}} [\log P(W|S, X) + \log P(X|S)] \quad (2)$$

The first term, $\log P(W|S, X)$, corresponds to the posterior probability of the word W estimated in our case through a speaker-dependent hybrid HMM/ANN with parameters θ_s as follows:

$$\log P(W|S, X) = \frac{1}{T} \sum_{t=1}^{T} \log p(q_k^t|x_t, \theta_s) \quad (3)$$

where q_k^t represents the "optimal" state q_k decoded at time t along the Viterbi path, and T the length of X after removing the decoded silence frames.

The second term, $\log P(X|S)$, corresponds to the likelihood of the observed data estimated by a text-independent GMM model with parameters λ_s:

$$\log P(X|S) = \frac{1}{T}\sum_{t=1}^{T} \log p(x_t|\lambda_s) \tag{4}$$

The $\log P(W|S,X)$ and $\log P(X|S)$ represent, respectively, the contribution of the speech and speaker recognition systems in the combined score. Using (2) for all registered speakers is time consuming. Therefore, we generate a list of N-best speakers according to the likelihood criterion (4) and then re-score this list using (2).

3 Database and Experiment setup

The experiments were done using the PolyVar database [6]. For the closed set experiments, a set of 19 speakers (12 males and 7 females) who were in more than 26 sessions are selected. Each session consists of one repetition of the same set of 17 words common for all speakers. For each speaker, the first 5 sessions are used as training (adaptation) data and an average of 19 sessions as test data, resulting in a total of 6430 test utterances. For the open set experiments, another set of 19 speakers with the same set of words are used as impostors. There are a total of 6452 impostor test utterances. For acoustic features, 12 MFCC coefficients with energy and their first derivatives were calculated every 10 ms over a 30 ms window.

We have also, used the PolyPhone database [6] to train a speaker-independent speech recognizer and a Gaussian Mixture model (GMM). They will be used only as an initial distribution for speaker adaptation.

- The SI speech recognizer is a hybrid HMM/MLP system[9], with a set of parameters Θ. This SI-MLP has 234 input units with 9 consecutive 26 dimensional acoustic vectors, 600 hidden units and 36 outputs.
- The GMM model with a set of parameters Λ is modeled by 240 (diagonal covariance) Gaussians and trained with an EM algorithm.

4 Speech and Speaker Recognition Approaches

In this work, we are interested in correctly recognizing both the pronounced word and the speaker identity for each test utterance. We have examined and compared three techniques. They used the same text-independent GMM based speaker identification subsystem. But, they employ different speech recognizers and different ways to integrate the speech and speaker recognizers.

4.1 Speaker identification subsystem

The speaker identification subsystem is a text-independent GMM based [7]. The parameters λ_s of the speaker-dependent GMM are derived (using the speaker's training data) by adapting mean parameters of mixtures components of the GMM model Λ. The adaptation is performed using MAP adaptation technique[8]. The correct speaker identification rate for the closed set is equal to 95.9%.

4.2 Baseline approach

Here the speech recognizer is a speaker-independent hybrid HMM/MLP system. The parameters θ of the MLP are derived by re-training all the parameters Θ of the SI-MLP trained on PolyPhone. The re-training is done using data (referred to as *world data*) from PolyVar provided by 56 speakers with the same set of 17 words. A cross-validation is used to avoid overtraining. The word recognition rate is equal to 97.2% and 96.8% for the closed set and open set applications, respectively. The recognition

of the pronounced word and the identification of the speaker are done independently. In the closed set application, this is done as follows:

$$\widehat{W} = \arg\max_{\{W\}} [\log P(W|\theta, X)] \tag{5}$$

$$\widehat{S} = \arg\max_{\{S\}} [\log P(X|\lambda_s)] \tag{6}$$

In open set application, the speaker is accepted if:

$$LLR(X) = \log P(X|\lambda_{\hat{s}}) - \log P(X|\lambda) \geq \delta \tag{7}$$

where $LLR(X)$ is the likelihood ratio, δ is a speaker and word independent threshold, $\lambda_{\hat{s}}$ is the GMM model of the most likely speaker $\widehat{S}$ according to (4), and λ is the *background* model where its parameters are derived from Λ using MAP adaptation and the *world data* set.

4.3 Sequential approach

Here, the speech recognition is performed using a speaker-dependent HMM/MLP. The parameters θ_s of the MLP are derived by re-training the parameters Θ of the SI-MLP, using speaker's training data. The most likely speaker determined by the speaker identification subsystem using (6) is used to select the SD-MLP for speech recognition. The speaker in both closed set and open set applications is identified as in the baseline approach using (6) and (7), respectively. The recognition of the pronounced word is performed as follows:

$$\widehat{W} = \arg\max_{\{W\}} [\log P(W|\theta_{\hat{s}}, X)] \tag{8}$$

where $\theta_{\hat{s}}$ is the set of parameters of the MLP associated with the most likely speaker $\widehat{S}$. With perfect recognition of the speaker identity, the word recognition rate is equal to 98.9%. The main advantage of this approach compared to the baseline is the gain we got in speech recognition performance. This gain should improve the performance of the simultaneous speech and speaker recognition. A system using this approach can be viewed as performing "speaker quantization" before speech recognition[1]. The speaker identification subsystem associates a new speaker to the most likely similar speaker in the enrolled speaker set.

4.4 Combined approach

The MLP adaptation for a specific speaker consists of shifting the boundaries between the phone classes without strongly affecting the posterior probabilities of the speech sounds of other speakers. This makes the estimated posterior probabilities more effective for speech recognition but less effective for speaker recognition [10][11]. Nevertheless, these posterior probabilities can be used to improve the speaker recognition performance if they are combined with more speaker specific information. In the closed set application, the recognition of both the pronounced word and the speaker identity is performed as follows:

$$(\widehat{W}, \widehat{S}) = \arg\max_{\{W,S\}} [\log P(W|\theta_s, X) + \log P(X|\lambda_s)] \tag{9}$$

Given both speaker identification and speaker-dependent speech recognition subsystems are trained with two different criteria, the posterior probability and the likelihood scores estimated by each subsystem might have some complementary (new) information that can be useful to improve the performance of each individual subsystem or the joint speech and speaker recognition. The criterion (9) should be done for every speaker and every word, making the computation requirement very costly (compared to the two previous approaches, we need an additional cost of $(N-1)$ times the cost of a speech recognition task, where N is the number of the enrolled speakers). To reduce this cost, we first generate a list

of the N-best candidates using text-independent speaker identification (6). Then, for each speaker S in the list, we use the SD-MLP θ_s for speech recognition. Finally, we re-score the N-best list according to the combined likelihood and posterior probability scores using (9). This procedure generates a new N-best list, where the most likely speaker is selected according to the following combined criterion:

$$\widehat{S} = \arg\max_{\{S\}} [\log P(W|\theta_s, X) + \log P(X|\lambda_s)] \tag{10}$$

For the open set application, the goal is to detect an impostor and reject him/her independently of what he/she pronounces. The criterion to accept a speaker is defined as follows:

$$[\log P(W|\theta_{\hat{s}}, X) + \log P(X|\lambda_{\hat{s}})] - \log P(X|\lambda) \geq \delta \tag{11}$$

which is equivalent to:

$$\log P(W|\theta_{\hat{s}}, X) + LLR(X) \geq \delta \tag{12}$$

In this work, we have tried a linear combination technique to combine posterior probabilities and likelihoods. The combined score in (9) and (12) are estimated, respectvelly, as follows:

$$(\widehat{W}, \widehat{S}) = \arg\max_{\{W,S\}} [\alpha_1 \log P(W|\theta_s, X) + \log P(X|\lambda_s)] \tag{13}$$

$$\alpha_2 \log P(W|\theta_{\hat{s}}, X) + LLR(X) \geq \delta \tag{14}$$

where α_1 and α_2 are determined *a posteriori* on the test set.

We can also use (10) as a speaker selection criterion to improve the performance of the "speaker quantizer".

5 Experiments and Results

The aim of these experiments is to evaluate, analyze and compare the effectiveness of the three different approaches described above in three different tasks, closed set speaker identification, open set speaker identification and speaker quantization. In the first two tasks, our interest is to improve the simultaneous speaker and speech recognition performance. In the results, this will be referred to as *overall recognition rate*. While in the third task, our aim is to improve the speaker-independent speech recognition performance in an open set application.

5.1 Closed set results

The results of the closed set experiments are shown in Table (1). It gives the performance of each approach in terms of speech, speaker and overall recognition. It is worth mentioning here, that the best *overall recognition* rate we can achieve will be equal to the lowest recognition rate given by speech and speaker subsystems. From these results, we can see that:

Approaches	Baseline	Sequential	Combined
Speech Reco.	97.2%	98.7%	98.7%
Speaker Reco.	95.9%	95.9%	96.8%
Overall Reco.	93.4%	95.1%	95.9%

Table 1: *Speech, speaker and overall recognition rates for different approaches*

1. The sequential approach gave better performance in terms of *overall recognition* rate than the baseline approach. This is due mainly to the improvement in the speech recognition rate. From the computational cost point of view, both approaches have the same cost. It is interesting to note here that the speech recognition rate in the sequential approach (98.7%) is almost equal to that obtained with perfect speaker identification (98.9%). This means, that the hybrid HMM/MLP model θ_s of the speaker S still recognizes correctly the pronounced word even if the speech segment comes from another mis-identified speaker.

2. Compared to the other approaches, the use of the combined posteriori probability and likelihood criterion (9) gave the best *overall recognition* (95.9%) rate. As a consequence, the speaker identification performance was also improved. This is because two speakers which are acoustically close in the speaker space are not necessary close in the speech space. So, selecting speakers based on one of these two components is not optimal. In Figure (1) we have plotted the variations of speech, speaker and overall recognition rates as a function of the size of the N-best candidates list. It shows that the most significant improvement is obtained by keeping the first two best likely speakers according to (6), and then use (9) for re-scoring. From the computational cost point of view, the combined approach needs only one more speech recognition step which depends on the size of the MLP and the length of the pronounced word.

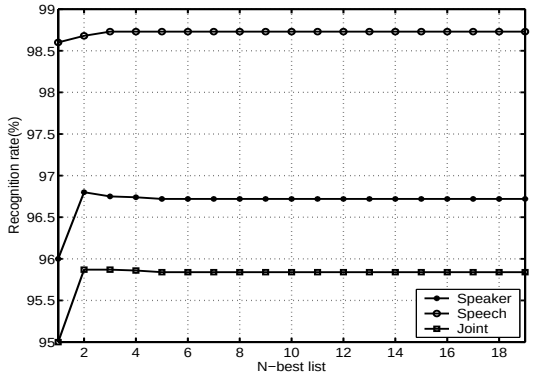

Figure 1: *Speaker, speech and overall recognition rates as a function of the size of the N-best candidates list*

5.2 Open set results

To evaluate our approach in a more practical application, open set experiments are conducted. The goal is to detect an unknown speaker (impostor) and reject him. Three types of errors are considered here [5], False acceptance (FA), false rejection (FR) and confusion acceptance (CA), that is, when an authorized speaker is accepted but confused with another speaker. We have plotted the variations of these errors as a function of a threshold for the sequential [1] (Figure (2)) and combined (Figure (3)) approaches.

The EERs (FA = FR) obtained by the sequential and combined approaches were equal to 14.5% and 13.1%, respectivelly. Moreover, the combined approach reduced the confusion acceptance errors. If we take into account only the true speakers that have been accepted, the *overall recognition* rates with the baseline, sequential and combined approaches were equal to 82%, 83.8% and 85.7%, respectivelly, confirming the tendency we have seen in closed set application.

[1] Both baseline and sequential approaches use the same speaker identification criterion (7) in an open set test

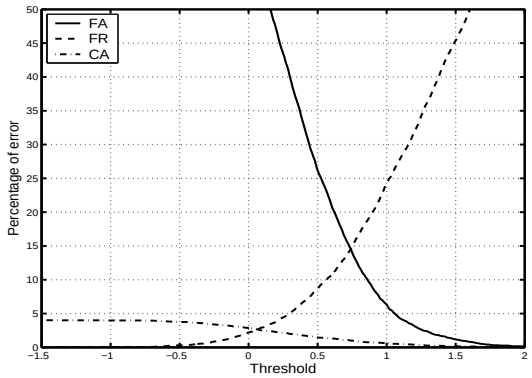

Figure 2: *False acceptance, false rejection and confusion acceptance variations as a function of the threshold for baseline and sequential approaches*

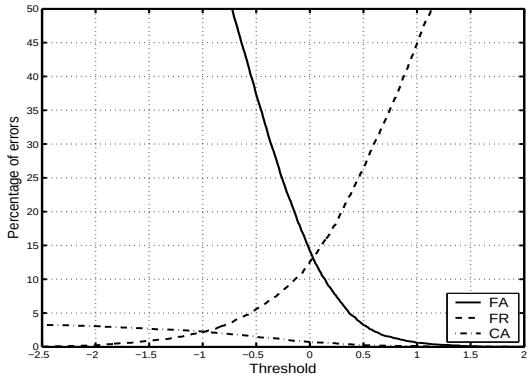

Figure 3: *False acceptance, false rejection and confusion acceptance variations as a function of the threshold for the combined approache*

5.3 Speaker quantizer results

In this experiment, we evaluate the use of the sequential and combined approaches to perform a speaker-independent speech recognition in open set application. This is done by selecting the enrolled speaker that is acoustically close to the test speaker and using the speech recognizer associated with the selected speaker to recognize the pronounced word. The main issue here is what will be the criterion to select the closet enrolled speaker? We have tested two criteria described in (6) and (10). Results of the speech recognition performance are shown in Table (2). For comparison purposes, the average performance of using each enrolled speaker is also reported (second column). We have used only impostor utterances (6452 utterances). As we can see, the use of the sequential criterion gave 8.4%

Approaches	Single speaker	Sequential	Combined
Speech Reco.	85.17%	92.3%	93.5%

Table 2: *Speaker quantizer performance for speech recognition*

relative improvement compared to the single speaker results. But the best improvement is achieved by the combined criterion (11% relative improvement). This is because, using (10) the selected reference speaker is acoustically close to the test speaker in the joint speech and speaker space.

6 Conclusion

In this paper, a probabilistic approach that maximizes simultaneous speech and speaker recognition performance is presented. It is based on the combination of posteriori probability estimated by a hybrid HMM/MLP system for isolated word recognition and likelihood estimated by a text-independent GMM model for speaker identification. We have evaluated and compared three approaches for closed set speaker identification, open set speaker identification and speaker quantization for speaker-independent speech recognition. In the three applications, results showed the effectiveness of the proposed approach.

7 Acknowledgment

The authors gratefully acknowledge the support of the Swiss National Science Foundation through the project "MULTI :2000-068231.02/1. This work was also carried on in the framework of the SNSF National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The authors would like to thank Hynek Hermansky for helpful comments on the paper and Joanne Moore for proofreading this paper.

References

[1] D. A. Reynolds and L. P. Heck, "Integration of Speaker and Speech recognition systems" *proceedings of ICASSP'91* pp. 869-872, 1991.

[2] L. P. Heck, "A Bayesian Framework for Optimizing the joint Probability of Speaker and Speech Recognition Hypotheses", *The Advent of Biometrics on the Internet*, COST 275 Workshop, Rome, Italy, 2002.

[3] Q. Li, B.-H. Juang, Q. Zhou, C.-H. Lee, "Automatic Verbal Information Verification for User Authentication", *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 5,2000.

[4] S. H. Maes, "Conversational Biometrics" *Proceedings of EUROSPEECH'99*, vol. 3, pp. 1219-1222, 1999.

[5] T. J. Hazen, D. A. Jones, A. Park, L. C. Kukolich and D. A. Reynolds, "Integration of speaker Recognition into Conversational Spoken Dialog systems" *Proceedings of EUROSPEECH'03* pp. 1961-1964.

[6] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, and P. Langlais, "Swiss French PolyPhone and PolyVar: telephone speech databases to model inter- and intra-speaker variability", *IDIAP Research Report*, IDIAP-RR-96-01, 1996.

[7] D. A. Reynolds, T. F. Quatieri and R.B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol.10, N 1-3, 2000, pp 19-41.

[8] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains", in *IEEE Transaction on Speech Audio Processing,* April 1994, Vol 2,pp. 291-298.

[9] S. Renals, N. Morgan, H. Bourlard, M. Cohen, H. Franco, "Connectionist probability estimators in HMM speech recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 1, Part II, 1994.

[10] D. Genoud, D. Ellis and N. Morgan, "Combined speech and speaker recognition with speaker-adapted connectionist models", *Proc. Auto. Speech recog. and Understanding Workshop,* keystone

[11] M. F. BenZeghiba and H. Bourlard, "User-Customized Password Speaker Verification based on HMM/ANN and GMM models", *Proceedings of ICSLP 2002*, pp 1325-1328, 2002.