IDIAP RESEARCH REPORT

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

On the combination of speech and speaker recognition

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

IDIAP–RR 03-19

April 10, 2003

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

IDIAP Research Report 03-19

On the combination of speech and speaker recognition

Mohamed Faouzi BenZeghiba Hervé Bourlard

April 10, 2003

Abstract. This paper investigates an approach that maximizes the joint posterior probability of the pronounced word and the speaker identity given the observed data. This probability can be expressed as a product of the posterior probability of the pronounced word estimated through an artificial neural network (ANN), and the likelihood of the data estimated through a Gaussian mixture model (GMM). We show that the posterior probabilities estimated through a speaker-dependent ANN, as usually done in the hybrid HMM/ANN systems, are reliable for speech recognition but they are less reliable for speaker recognition. To alleviate this problem, we thus study how this posterior probability can be combined with the likelihood derived from a speaker-dependent GMM model to improve the speaker recognition performance. We thus end up with a joint model that can be used for text-dependent speaker identification and for speech recognition (and mutually benefiting from each other).

1 Introduction

Speech recognition aims to extract the lexical information from a speech signal. Speaker recognition aims to recognize (identify or verify) the speaker's identity from a signal with knowledge of the lexical content (text-dependent) or without (text-independent). Both tasks take the same signal as input but for two different purposes. As a result, speech recognizers are designed to ignore or remove the unwanted information in the signal that may be useful for speaker recognition and vice versa [1]. However, the scores estimated by both recognizers might contain complementary information that can be combined to improve the performance of each of the systems used independently.

Joint speech and speaker recognition is important for many applications, such as information retrieval from audio data, interaction with an automated voice system and verbal information verification [2], and in general, in any application where we are interested in *who is speaking and what was said*.

In [1], there are two applications presented. The first one is named-based identity claim recognition, where the speaker recognition subsystem is used to help guide the speech recognition search for the identity claim. The second one is speaker and speech recognition of co-channel speech when more than one speaker are speaking at the same time. In this application, the speaker recognition score is used as a reliability measure.

There are some applications where all the registered speakers share the same set of commands (words) and where each command is associated with a specific service. To make the use of the application more friendly, a convenient way is to let customers access the services by just pronouncing the command associated with the service. The task of the system is then to simultaneously recognize the command and to identify the speaker. A typical system for this application is to use a speaker-independent speech recognition to recognize the command and a speaker recognition system to identify the speaker identity. The overall performance of the system depends on the individual performance of the speech and the speaker recognition subsystems.

In this paper, we present a new approach to combine the speech and the speaker recognition scores. More precisely, we use the speech recognition score to improve the speaker recognition performance. The speech recognition system is based on the hybrid HMM/ANN systems [3], where an Artificial Neural Network (ANN) is used to estimate Hidden Markov Model (HMM) state emission posterior probabilities. These systems are suitable for the application we are interested in, as it has been found that the estimated posterior probabilities are a good confidence measure to indicate how good the word model matches the data [4, 5]. The speaker recognition subsystem is based on a state-of-the art Gaussian mixture model (GMM) approach [6].

The rest of this paper is organized as follows: In the next session, we present the formulation of the problem. Section 3 describes the database used for evaluation. In Section 4, we describe the two approaches that are studied andSection 5 describes the experiments and discusses the obtained results.

2 Formulation

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

$$\begin{aligned}(\hat{W},\hat{S}) &= \arg\max_{\{W,S\}} P(W,S|X) \\ &= \arg\max_{\{W,S\}} \left[P(W|S,X)\cdot P(S|X)\right] \qquad (1)\end{aligned}$$

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

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

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

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

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, $\log P(X|S)$, represents the contribution of the speaker recognizer part and corresponds to the likelihood of the observed data estimated by a speaker dependent GMM model with parameters λ_s:

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

It can be observed that using (2) for all registered speakers is time consuming. To overcome this problem, we decided to generate a list of N-best speakers according to the likelihood criterion (4) and then rescore this list by combining the speech recognition score (3).

3 Database and Experiment setup

The experiments were carried out using PolyVar database [9]. A set of 19 speakers (12 males and 7 females) who have 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. Another data (referred to as *world data*) from a set of 56 speakers with the same set of words are used to train a speaker-independent MLP and speaker-independent GMM models. For acoustic features, 12 MFCC coefficients with energy complemented by their first derivatives were calculated every 10 ms over 30 ms window, resulting in 26 coefficients.

4 Approaches and performance measurement

4.1 Baseline system

A typical system for our application is to use a speaker-independent speech recognizer to recognize the pronounced word and a speaker recognition system to identify the speaker identity.

- The speech recognizer is a speaker-independent hybrid HMM/ANN system. The ANN is a multi-layer perceptron (MLP) with parameter θ trained using *world data*. A cross-validation technique is used to avoid overtraining. This SI-MLP consists of 130 input units with 5 consecutive frames, 200 hidden units and 28 outputs, where each output belongs to a specific phoneme in the dictionary of the application.

- For each registered speaker, a speaker-dependent GMM model λ_s is created by adapting mean parameters of mixtures components of a speaker-independent GMM (SI-GMM) trained with *world data*. This SI-GMM consists of 200 mixtures components

The goal of the system is thus to recognize correctly both the pronounced word and the speaker identity. For the baseline system, this will be referred to as the *independent overall recognition* [1] and it is done as follows:

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

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

where $\log P(W|\theta, X)$ and $\log P(X|\lambda_s)$ are estimated respectively according to (3) and (4). To improve the overall performance, we have to improve the performance of both speech and speaker recognition subsystems.

4.2 Combined (proposed) system

To improve the speech recognition performance, we have used a speaker-dependent speech recognizer. So, in the proposed approach, each speaker is modeled by two subsystems, a speaker-dependent GMM model λ_s, as used in the baseline system, and a speaker-dependent MLP θ_s created by re-training (using the speaker training data) the speaker-independent MLP θ used in the baseline system. The segmentation is obtained by force aligning the utterance on the hypothesized word HMM model using local posterior probabilities estimated by the SI-MLP. A cross-validation technique is used to avoid overtraining. The recognition of both the pronounced word and the speaker identity, which will be referred to as the *joint overall recognition* is then based on the following criterion:

$$(\hat{W}, \hat{S}) = \arg\max_{\{W,S\}} [\log P(W|\theta_s, X) + \log P(X|\lambda_s)] \tag{7}$$

This should be estimated for every possible word and every possible speaker. To reduce the computational cost, we, first generate a list of the N-best speakers according to the likelihood criterion (4). Then, for each speaker S in the list, we perform a speech recognition step using the associated SD-MLP θ_s to estimate the phone emission posterior probabilities which are then used in Viterbi decoding to recognize the pronounced word yielding the highest time normalized posterior probability (3). Finally, we rescored the list according to the combined likelihood and posterior probability scores, resulting in the following speaker recognition criterion:

$$\hat{S} = \arg\max_{\{S\}} [\log P(W|\theta_s, X) + \log P(X|\lambda_s)] \tag{8}$$

4.3 Connectionist system

We can also try to use the SD-MLP θ_s for both speech and speaker recognition. Formally, this is can be expressed as:

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

where $\log P(W|\theta_s, X)$ is estimated using (3). The SI-MLP is trained to discriminate between the phone classes. The adaptation for a specific speaker consists in shifting the boundaries between these classes without any direct influence on the posterior probabilities of the speech sounds of other speakers [8]. This makes the estimated posterior probability more effective for speech recognition but less effective for speaker recognition [7]. Nevertheless, these posterior probabilities can be used to improve the speaker recognition performance if they are combined with more speaker specific score (likelihood)(4). Recognition of both the pronounced word and the speaker identity using(9) will be referred to as the *connectionist overall recognition.*

[1] Because both the pronounced word and the speaker identity are recognized independently

5 Experiments and Results

The aim of the experiments is to show how the speech recognition posterior probability based score can be used to improve the speaker recognition likelihood based performance and then the *joint overall recognition* performance. The test is performed in a *closed-set*. In the context of this application, this means that the pronounced word and the unknown speaker should be assigned to the word and speaker whose models match best the unknown speaker's utterance test. We have conducted two sets of experiments. The first set concerns the evaluation of the baseline and the connectionist systems. The second set concerns the evaluation of the combined system with reference the results obtained by the baseline system.

5.1 Baseline and connectionist systems evaluation

	Baseline	*Connectionist*
Speech recognition	96.27%	96.11%
Speaker recognition	96.03%	52.17%
Independent/connectionist overall recognition	92.45%	51.80%

Table 1: *Speech, speaker and overall recognition rates with the baseline system ((5) and (6)) and the connectionist approach ((9)).*

Table 1 shows the results obtained by the baseline and the connectionist systems. The first two lines (speech and speaker) correspond respectively, to the speech and speaker recognition rates. For the baseline system, these are can be obtained by independently using (5) for speech recognition and (6) for speaker recognition. For the connectionist system, we look at the recognized word (speech) and the recognized identity (speaker) obtained by (9). The third line (*independent* for baseline system and *connectionist* for connectionist system) gives the correct overall recognition rate simply computed by counting the number of times the correct speaker and the correct word occurred simultaneously.

Concerning the connectionist approach, one can observe that contrary to the performance of the speaker recognition where the result is worse (52.17%), the result of the speech recognition is satisfactory (96.11%). This means that a speaker model (SD-MLP) θ_s still recognizes correctly the pronounced word even if the speech comes from a different speaker. As explained in (Section 4.3), it shows that the posterior probabilities used alone are much better for speech recognition than for speaker recognition.

It is obvious that the best *independent/connectionist overall recognition* rates we can achieve will be equal to the lowest of the speech and speaker recognition rates (in our case, the speaker recognition rate). As there are some accesses, where the speaker is correctly identified and the pronounced word is wrongly recognized, this makes the speaker recognition rate better than the *overall recognition* rate in both systems.

5.2 Combined system evaluation

By using the *combined system*, our aim is to obtain a *joint overall recognition* rate better than the *independent overall recognition* rate obtained by the baseline system. This is can be achieved by taking the advantage of the use of a speaker-dependent speech recognizer to:

1. Improve the speech recognition rate.
2. Improve the speaker recognition rate by using (8).

The size of the N-best list should be chosen as a tradeoff between the *joint overall recognition* performance (7) and the computational cost. As we can expect from the results above, using (8) for speaker recognition, the likelihood estimated by the SD-GMM λ_s will have more contribution than the posterior probability. So, if the correct speaker does not appear in the first few positions in the N-best list, the use of (8) to recognize the speaker is not going to help much.

Table 2 shows the results obtained by the combined system, for different sizes of the N-best list.

	Combined system		
N-best list size (N)	N=1	N=2	N=3
Speech recognition	97.91%	97.91%	97.91%
Speaker recognition	96.03%	96.51%	96.55%
Joint overall reco.	94.10%	94.53%	94.58%

Table 2: *Speech, speaker and the joint overall recognition rates for different sizes of the N-best list: The speech recognition rate is obtained by (5), where a SD-MLP θ_s was used instead of the SI-MLP θ. The speaker recognition and the joint overall recognition rates are obtained respectively, by using (8) and (7).*

It can be observed that the combined system performs always better (in terms of the *overall recognition* rate) than the baseline system (see Table 1, first column).

For N equal 1, there is no difference between the combined system and the baseline system from the computational cost point of view. Both systems have the same number of parameters. The only difference is that for speech recognition, in the proposed approach, we have used the SD-MLP θ_s associated with the identified speaker (whose GMM model matches best with the observed data) instead of using the SI-MLP θ for posterior probabilities estimation. For this case,we can see that the *joint overall recognition* rate (94.10%) is significantly better than the *independent overall recognition* rate (92.45%) obtained by the baseline system. This is mainly due to the use of speaker-dependent speech recognizer, which gave a 97.91% word recognition rate.

Furthermore, as we have seen in the introduction, the information extracted from the signal can be different depending on the task. For example, for speech recognition, this information can be some *meaningful* sounds like phonemes, while for speaker recognition, this information is more related to the speaker characteristics. Having said that, two speakers which are close in the likelihood domain are not necessary close in the posterior probability domain, and vice-versa. Sometimes, it happens that the data matches better with another speaker's model than the correct one. In such a case, the correct speaker identity will appear in the second position (or more) in the list. By using (6) as a criterion for speaker recognition, the correct speaker will be misidentified. But, by using the combined score (8), the correct speaker can appear in the first position. As we can see from the Table 2, for N equal 2, the speaker recognition rate improved from 96.03% to 96.51%. As a consequence, the *joint overall recognition* performance is also improved (94.53% compared to 94.10%). From the computational cost point of view and compared to the case where N equal 1, we need one more speech recognition step. The cost of this step is dependent on the number of parameters of the SD-MLP. The improvement in the *joint overall recognition* becomes more insignificant as the size of the list increases, making the choice of N equal 2 a good tradeoff between the *joint overall recognition* and the computational cost.

6 Conclusion

In this paper, a new approach that maximize the joint posterior probability of speech and speaker is presented. This probability can be expressed as a product of the posterior probability of the pronounced word given the data and the likelihood of the data given a speaker model. A comparative study is carried out with a baseline system where both the pronounced word and the speaker identity are recognized independently. We have studied how the posterior probability estimated through a speaker-dependent MLP and the likelihood estimated through a speaker-dependent GMM model can be mutually beneficial. So, the joint speech and speaker recognition performance can be significantly improved compared to the baseline system.

7 Acknowledgment

Mohamed F. BenZeghiba is supported by the Swiss National Science Foundation thought the project "SV-UCP: Speaker Verification based on User-Customized password" (2000-063721.00-1).

References

[1] 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.

[2] 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.

[3] H. Bourlard and N. Morgan, "Connectionist Speech Recognition: A hybrid approach", Kluwer Academic Publisher,1994.

[4] G. Williams and S. Renals, "Confidence Measures for hybrid HMM/ANN speech recognition", *Proceedings of Eurospeech'97*, pp. 1955-1958, 1997.

[5] G. Bernardis and H. Bourlard, "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN speech recognition systems", *Proc. of Intl. Conf. on Spoken Language Processing* (Sydney), pp. 775-779, 1998.

[6] 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.

[7] M. F. BenZeghiba and H. Bourlard, "User-Customized Password Speaker Verification based on HMM/ANN and GMM models", *Proceedings of ICSLP 2001*, pp 1325-1328, 2002.

[8] D. Genoud, D. Ellis and N. Morgan, "Combined Speech and Speaker Recognition with Speaker-Adapted Connectionist Models", *Proceedings Auto. Speech Recognition and Understanding Workshop*, Keystone, 2001.

[9] 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.