Joint Speaker Identification and Speech Recognition for Speech Controlled Applications in an Automotive Environment

T. Herbig[1,2], F. Gerl[2]

[1] University of Ulm, Dept. of Information Technology, Ulm, Germany, Email: tobias.herbig@harman.com

[2] Harman Becker Automotive Systems, Speech Dialog Systems, Ulm, Germany, Email: franz.gerl@harman.com

Introduction

In recent years speech recognition and speaker identification have increasingly obtained attention for a variety of speech controlled applications such as navigation and handsfree phones in an in-car application.

Speaker dependencies can severely degrade the performance of speech recognition due to the mismatch between the speech recognizer's training and the speech utterances during runtime. There are two possibilities to deal with the speaker characteristics: Normalization techniques during the feature extraction process can assure an increased robustness of speech controlled systems against speaker-dependencies. Another way is to include the speaker's characteristics in the speech recognition by applying speaker adaptation techniques to achieve a better statistical representation of the user's speech.

There exists a variety of approaches for adapting the statistical model of a speech recognizer so that the characteristics of a particular user are better represented. They mainly differ in the number of parameters which have to be estimated. A higher number of parameters generally allows better adaptation resulting in higher speech recognition rates. The amount of speech required for adaptation however closely corresponds to the number of parameters. Thus truly speaker independent systems usually adapt using only very few utterances to avoid misadaptations following speaker changes.

For a system that is used only by a limited number of users as in e.g. in-car applications this problem can be circumvented if one tracks the identity of the speakers. This motivated us to introduce a reliable speaker identification which enables the speech recognizer to create robust speaker dependent models. Speaker identification using voice only is a well known problem with many approaches described in the literature. Even though speech and speaker recognition commonly apply the same statistical methods to similar features tackling them together has rather been the exception so far [1].

The goal of this paper is to introduce an integrated approach for speech and speaker recognition. A speaker identification device is integrated into an existing speech recognizer. The interaction of speech and speaker recognition enhances both recognition rates concerning speech and speaker identity as speech recognition exploits the knowledge of the speaker and vice versa. The speaker identification supports the speech recognizer to build up speaker-dependent models for speech recognition and speaker identification by unsupervised online speaker adaptation starting from a speaker-independent codebook. Thus speaker adaptation is enabled to adapt to different speakers which results in an optimal long-term adaptation.

Unlike existing speaker identification systems this approach requires only a short training phase for new speakers. During the speech decoding process the identity of the current speaker is estimated and the speaker dependent codebook is further adapted after each identified utterance. Thus the speech controlled system has to handle speaker changes, speaker identification and speaker adaptation in an unsupervised manner. Furthermore the system accounts for different training levels of the statistical models by providing a smooth transition between short-term and long-term speaker adaptation depending on the amount of speaker-dependent utterances.

In the following a brief introduction to speaker identification with *Gaussian Mixture Models* (GMM) and speech recognition based on semi-continuous *Hidden Markov Models* (HMM) is given. Fast and long-term speaker adaptation are briefly described and the proposed combination is presented. Then the new joint speech and speaker recognition system is described. The system is evaluated in an realistic in-car environment for a limited number of enrolled speakers. Speech and speaker recognition rates are presented for a subset of the Speecon database.

Speaker Identification

Gaussian Mixture Models have been established as the dominating state-of-the-art statistical model for speaker identification due to their flexibility in representing arbitrary probablitity density functions [2]. A GMM comprises a set of N multivariate Gaussian distributions which are entirely defined by their mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. j and k denote the class and the speaker index. The weighting factor w represents the impact of the Gaussian distribution compared to the remaining distributions and is derived during the GMM training. The convex combination of all classes results in the likelihood function $p(\mathbf{x}_t|\boldsymbol{\lambda}^k)$ of the GMM for a particular feature vector $\mathbf{x}$ at time instance t. $\boldsymbol{\lambda}$ comprises all weights, mean vectors and covariance matrices.

$$p(\mathbf{x}_t|\boldsymbol{\lambda}^k) = \sum_{j=0}^{N-1} w_j^k \cdot \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_j^k, \mathbf{\Sigma}_j^k) \quad (1)$$

The training of a GMM can be realized by the *expectation maximization* (EM) or the k-means algorithm, for example.

In the following an utterance $\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_T]$ is considered as a sequence of feature vectors $\mathbf{x}_t$. Under the *iid* (independent and identically distributed) assumption the logarithmic likelihood for a particular speaker k can be derived from equation (1).

$$\log(p(\mathbf{X}|\boldsymbol{\lambda}^k)) = \sum_{t=0}^{T} \log(\sum_{j=0}^{N-1} w_j^k \cdot \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_j^k, \boldsymbol{\Sigma}_j^k)) \quad (2)$$

In state-of-the-art identification systems the speaker with the highest likelihood $p(\mathbf{X}|\boldsymbol{\lambda}^k)$ is detected as the current speaker [3].

Speech Recognition

A speech recognizer based on semi-continuous HMMs can be divided into three sections as depicted in Figure 1.

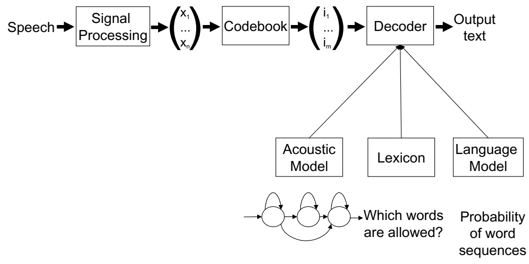

Figure 1: Speech recognition using semi-continuous HMMs

First, a pre-processing and a feature extraction are applied to the sampled and digitized speech signal. In the preprocessing step the speech signal is enhanced by common noise reduction algorithms, e.g. a standard Wiener filter. In each time step the feature extraction maps the speech signal into a vector representation by extracting the significant characteristics of the speech.

For many speech recognizers *Mel Frequency Cepstral Coefficients* (MFCCs), the first and second derivatives are computed. The first element of the MFCC vector is usually replaced by a normalized logarithmic energy estimate. The remaining MFCCs are mean subtracted to reduce the influence of the channel and the mircophone characteristic.

Second, the feature vector is compared with a speaker independent or a speaker dependent codebook. The codebook consists of several multivariate Gaussians representing the emission probability of the underlying Markov model. Whereas the hidden Markov process provides a statistical model of the speech context, the codebook represents the statistical variations in the pronunciation which are usually highly speaker dependent. This comparison is done by determining a normalized likelihood for each class.

Third, this match between feature vector and codebook is used for the speech decoding step comprising an acoustical model, the lexicon and the language model. The acoustical model is realized as a Markov model of first order as the hidden statistical process of the HMM.

The result is a transcription of the spoken utterance.

Speaker Adaptation

The purpose of speaker adaptation as described here is the modification of the codebook to obtain a better match between the statistical model and the real characteristics of the observed utterance. In unsupervised speaker adaptation the codebook is modified so that the likelihood of the recognized word sequence is increased.

Fast adaptation algorithms like the eigenvoice adaptation [4] are capable to extract the significant speaker characteristics even in a sparse data scenario. The eigenvoices take prior knowledge of the statistical relations among the Gaussians into account so that even with a small number of parameters the codebook can be adapted efficiently. In an offline training step the main translations of the Gaussian centers are learned which are observed for long-term adaptation. The adaptation during runtime is restricted to these translations so that only the step size of each translation has to be estimated. The eigenvoice approach permits to adapt all classes of a codebook even if no or only sparse data are available for particular classes. The small number of parameters permit robust estimates and reduces the risk of overfitting.

Long-term adaptation algorithms such as the *Maximum A Posteriori* (MAP) algorithm [5] are well suited if sufficient utterances of a given speaker are available. The MAP algorithm adapts all classes separately and does not account for inter-class dependencies. The MAP algorithm performs one step of the EM algorithm to compute a new set of mean vectors and provides an interpolation between the prior mean values and the new estimates. The number of feature vectors softly assigned to a particular class controls the interpolation. If the new estimate is calculated on sparse data the interpolation tends to keep the prior mean vectors. If the number increases the resulting mean vector converges towards the *Maximum Likelihood* (ML) estimate.

Here only the mean vectors of the codebook are modified as adapting the mean vectors usually achieves the highest improvement of the speech recognition rate.

The eigenvoice approach is selected for fast adaptation in the start phase of the system when new speakers are enrolled with sparse training data. The eigenvoices represent the most likely directions in which the mean vectors are adapted as seen in the training. The weight of the eigenvoices contains the speaker dependencies. Here only the first 10 eigenvoices representing the largest eigenvalues are considered.

The presented algorithm applies the MAP adaptation to the speaker dependent codebooks as soon as the speaker identification assigns a sufficient number of utterances to a particular speaker. This guarantees that the codebook learns even those speaker characteristics that the eigenvoice approach cannot capture.

The transition between the eigenvoice and the MAP adaptation is realized as a convex combination depending on the number of adaptation data. In [6] a similar combination of eigenvoices and MAP adaptation is reported. Whereas [6] applies a matrix multiplication this approach only requires a scalar product which reduces the computational load significantly.

Joint Speaker Identification and Speech Recognition

The new approach for a unified framework for speech recognition and speaker identification is based on the observation that both applications use similar statistical models and even similar input features. As the codebook of a speech recognizer and the GMM of the speaker identification apply a set of multivariate Gaussian distributions, the new approach uses speaker dependent codebooks simultanously for speech decoding and speaker identification. Thus the complexity of two parallel independent systems is avoided.

The approach is realized by adding N speaker dependent codebooks in parallel to the speaker independent codebook which is also called standard codebook. The parallel computation of all enrolled speakers enables the speech recognizer to react rapidly to speaker changes.

Figure 2 describes the general configration of this unified approach.

The feature extraction is controlled by the speaker identification. Features such as energy and a mean normalization contain relevant speaker information and have to be computed for each speaker separately. Since N parallel feature extractions would increase the complexity substantially only the parameters of the previously detected speaker are considered. This implies the assumption that the speaker change rate is rather low which is valid for the applications considered.

Experimental investigations support that this restriction does not degrade the speaker change detection rate significantly but reduces the complexity of the system remarkably.

The feature vector is compared with the speaker independent codebook and a subset of Gaussians S with highest likelihood is extracted. Afterwards, the speaker dependent codebooks are investigated on the identical subset of Gaussians to reduce the computational load. In the final step the likelihood for each codebook is updated in each time step.

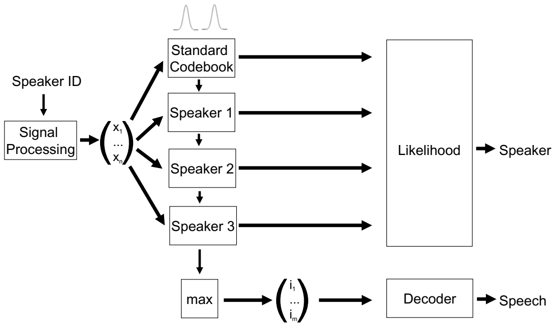

Figure 2: Joint speaker identification and speech recognition

$$\log(p(\mathbf{X}|\boldsymbol{\lambda}^k)) = \sum_{t=0}^{T} \log(\sum_{j\in S} \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_j^k, \boldsymbol{\Sigma}_j^k)) \qquad (3)$$

In contrast to (1) the sum only accounts for the Gaussians within the subset. The weights are omitted which is equal to a GMM with equally distributed weights. This is due to the observation that the likelihood of the Gaussians are sufficiently discriminative for speaker identification.

The evolution of the likelihood in (3) allows a fast decision during runtime and enables the speech recognizer to realize a speaker change by selecting another codebook for speech decoding.

The final likelihood score after each utterance determines the speaker of the utterance so that the speech recognizer can further adapt the corresponding codebook.

Results

A subset of the US-Speecon database [7] is used for the evaluation of the joint speech and speaker recognition. The subset comprises 73 speakers (50 male and 23 female speakers) recorded in a real automotive environment.

The sampling rate is 11025 Hz and only the AKG microphone recordings are investigated in the evaluation.

Colloquial utterances with more than 4 words and mispronunciations are removed to obtain a realistic command and control application. Digits and spelling loops are kept irrespective of their length.

This results in at least 250 utterances per speaker which are ordered in the sequence of the recording session.

The speech recognizer applies grammars for digit and spelling loops as well as for dedicated numbers. Finally a grammar that contains all remaining utterances is built.

The evaluation is computed on 60 sets of 5 enrolled speakers whereas the composition of the speaker groups is chosen randomly. The composition of female and male speakers in a group does not need to be balanced.

At the beginning of each speaker set an enrollment takes place consisting of 10 utterances for each speaker. In this paper only an in-set scenario is investigated in the sense that the system has to recognize a speaker from a list of enrolled speakers. Thus the first two utterances of each new speaker are indicated during the enrollment. After the first two utterances the system has to identify the new speaker whose speaker dependent codebook has to compete against the existing speaker models.

After the enrollment the speakers of a set are investigated in blocks of at least five utterances. The number of utterances within a block and the order of the speakers in each set are chosen randomly. For each set all 250 utterances of each speaker are used only once and a overall speaker change rate of 10% is enforced. Thus every set provides 1250 utterances originating from 5 speakers which results in 75000 utterances in total.

No external information concerning speaker changes and the speaker identity are given except for the first two utterances of each speaker.

The setup is chosen to guarantee independence of the speaker set composition. Furthermore this configuration investigates the robustness of the start phase, error propagation and the long-term speaker adaptation as well. Both the speaker recognition rate and the *word accuracy* (WA) of the speech recognizer are given in Table 1.

The baseline comprises the speaker independent speech recognizer which represents the lower limit of speech recognition rate.

In case of nonexisting speaker identification only a limited speaker adaptation can be applied as speaker changes cannot be detected. This would degrade the performance of the adapted codebooks. Therefore a time decay has to be introduced which limits the memory of the adaptation and leads to a non-optimal solution. In Table 1 this approach is denoted as short-term adaptation.

If perfect knowledge of the speaker identity is assumed the upper limit for the enhancement of speech recognition by unsupervised long-term adaptation is obtained.

Both the normalized energy and the mean subtraction are speaker dependent and are controlled by the speaker identification.

Description	Recognition rates [%]	
	WA	Speaker ID
Baseline	85.23	-
Short-term adaptation	86.13	-
Joint speech and speaker recognition	88.06	93.53
Perfect speaker ID	88.65	100

Table 1: Speech and speaker recognition rates

Table 1 directly shows that the speech recognition rate can be increased significantly if speaker adaptation is applied. The joint approach achieves 23.7% relative improvement of the WA compared to the baseline and 16.2% relative improvement compared to the short-term adaptation. The speaker identification of 93.53% verifies the robustness of the algorithm against error propagation caused by misclassification and continuous adaptation.

References

[1] Jesper O. Olsen: *Speaker Verification based on Phonetic Decision Making*, In EUROSPEECH-1997, pp. 1375-1378.

[2] Christopher M. Bishop: *Pattern Recognition and Machine Learning*, Springer, Berlin, 2008.

[3] Douglas A. Reynolds, Richard C. Rose: *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*, IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, 1995.

[4] R. Kuhn, P. Nguyen, J.-C. Jungua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, M. Contolini: *Eigenvoices for Speaker Adaptation*, Proc. of International Conference on Spoken Language Processing, pp. 1771-1774, 1998.

[5] Jean-Luc Gauvain, Chin-Hui Lee: *Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations for Markov Chains*, IEEE Transactions on Speech and Audio, 1994.

[6] Henrik Botterweck: *Anisotropic MAP defined by Eigenvoices For Large Vocabulary Continuous Speech Recognition*, Proc. Acoustics, Speech and Signal Processing, ICASSP 2001.

[7] D. J. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, A. Kiessling: *SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation*, Proceedings LREC'2002, pp. 329-333, 2002.