Noise-Robust Speaker Identification in Cars

Abhijatha Banashankarappa[1], Steffen Kortlang[1], Stephan Werner[2], Thomas Rohdenburg[1]
[1]IAV GmbH, Rockwellstrasse 5, 38518 Gifhorn, E-Mail: thomas.rohdenburg@iav.de
[2]TU Ilmenau, Electronic Media Technology, 98693 Ilmenau.

Abstract

Speech is one of the important communication tools between the human and the machine within a car. Besides voice recognition, the speaker identity is an important information extractable from the speech signal. By detecting the speakers, the infotainment system may automatically be personalized according to their preferences and it may be utilized for speaker dependent speech recognition. The focus of this work lies in audio-based speaker identification in cars using the existing hands-free system. Many different features like Gammatone and Mel Frequency Cepstral Coefficients, Linear Predictive Cepstral Coefficients, Linear Predictive Coefficients and corresponding delta and delta-delta features are extracted from the speech. Linear Discriminant Analysis is used to reduce the dimensionality of the features. Gaussian Mixture Models are used as the classifier. The system is implemented to operate in real-time.

The speech data is collected from 11 different speakers inside the car, using the built-in hands-free microphones. The actual driving noise measured at 60 & 120 km/h is added to the signals. A classification accuracy of 99.1%, 99.8% and 98.3% is achieved when the car is in idle condition, at 60 km/h and at 120 km/h, respectively.

Introduction

Speaker identification is a task of identifying who is speaking in a certain environment. There are two categories in speaker identification. *Text-dependent identification,* where the system is trained using a certain set of words and *text-independent identification*, where the system identifies the speaker from any words spoken. The latter is the focus of this work.

The infotainment system is an integral part of a car. The identification of the user inside helps in automatic personalization of the infotainment system according the users preferences. Reynolds et al. (1994) first introduced a text-independent speaker identification system using Gaussian Mixture Models (GMM) under normal acoustic conditions [7]. However, reverberation and running noises degrades the quality of the speech inside the car. This work evaluates the performance of the identification system under the realistic acoustic conditions within a car by extracting relevant features.

Measurement set-up

For this task, the speech data is collected using the recording set up built inside the car as shown in Figure 1. Two omnidirectional built-in microphones, near the central mirror, are used to record the speech from the front seat positions. Two omnidirectional microphones were installed at the back to record the speech signals from the rear seat positions. The distance between the two pair of microphones are maintained the same throughout the recordings (10 cms). The front microphones were situated in the intended positions and the rear microphones were manually installed. Only the speech recordings from the front-left position are considered for this work.

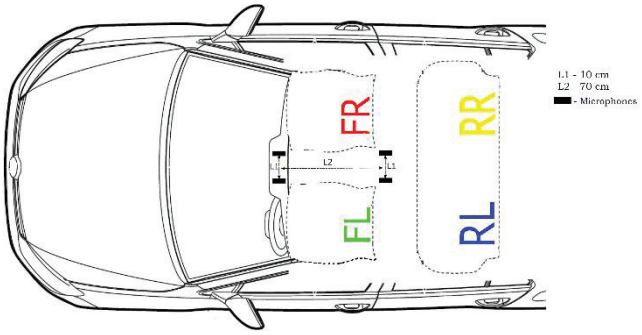

Figure 1: A schematic diagram of the measurement set-up inside the car.

Feature Extraction

As speaker identification has been the topic of interest from over four decades, many different features were proposed for this task [1]. Out of many, this work uses four different features that have proved effective for speaker identification. Mel Frequency Cepstral Coefficients (MFCC) are selected, as they have proven effective for speaker identification in many researches [3, 4]. Since they tend to fail under noisy conditions, Gammatone Frequency Cepstral Coefficients (GFCC) are selected, as they were shown to be more noise robust [2]. Linear Predictive Coefficients (LPC) are commonly used for spectral envelope representation of speech signals and Linear Predictive Cepstral Coefficients (LPCC) represents LPC in Cepstral domain [5]. Additionally, differential and acceleration coefficients are also extracted from the speech signals as they give a good representation of changes in the features with time.

Dimension Reduction

The feature set constitutes of 12 features, each with 12 dimensions, resulting in 144 dimensions. This is a huge set of data to represent a single audio frame, which might contain some redundant information. Therefore, there is a need to use a dimension reduction algorithm that selects the important features. Linear Discriminant Analysis (LDA) performs this task by projecting the feature set onto a hyper-plane. This method considers both the class and the feature-set to obtain a higher discriminating power.

Equation (1) represents the transformation matrix of LDA, known as Fischer's criterion.

$$S_W^{-1} S_B W = \lambda W \quad (1)$$

where, W is the transformation matrix, S_B is the between-class variance, S_W is the within-class variance and λ represents the eigenvalues of W. The solution for this problem is obtained by calculating the eigenvalues. The robustness of eigenvalues represents the ability to discriminate between different classes [6].

Gaussian Mixture Models

Gaussian mixture models have been widely used in the field of speech processing, for speaker identification, speech recognition and denoising of speech signals [1, 7].

A Gaussian mixture model is developed for speakers by modeling the distribution of their feature-set. For a feature-set x, the mixture densities are defined as [8]

$$p(x|\lambda) = \sum_{m=1}^{M} p_m d_m(x), \quad (2)$$

where, M is the number of mixture components, $d_m(x)$ is the Gaussian density, parameterized by mean vector μ_m and covariance matrix $\sum_m$, and p_m is the weight of the m^{th} component, satisfying the condition $\sum_{m=1}^{M} p_m = 1$.

The expectation maximization algorithm is used to estimate the maximum-likelihood parameters.

$$\hat{K} = \underset{1 \le k \le K}{\operatorname{argmax}} \sum_{l=1}^{L} \log p(x_l|\lambda_k), \quad (3)$$

In Equation 3, the speaker model λ_k that generates the highest value is selected as the identified speaker $\hat{K}$.

Data Collection & Test Cases

To train the classifier, the speech data is collected from 11 different speakers for a duration of 5 minutes using the built-in hands-free microphones inside the car. One of the noise sources that affects the speech in car is running noise, which constitutes of the noise generated by the engine, tire and wind [9]. Hence, actual car noise recorded at 60 km/h & 120 km/h is added to the speech signals. The performance of the system is evaluated for three test cases.

- When the car engine is on (idle)
- When the car is moving at 60 km/h
- When the car is moving at 120 km/h

When dealing with speech in the presence of background noise, it is important to consider the 'Lombard effect', which states that the humans have a tendency to increase their speech level in the presence of a background noise [10]. The approximate relation between the background noise level and the increase in the speech level is given from ITU-T recommendations and it is defined as [11]

$$I(N) = \begin{cases} 0, & N < 50 \\ 0.3\,(N - 50), & 50 \le N < 77, \\ 8.0, & N \ge 77 \end{cases} \quad (4)$$

where, I is the dB increase in speech level in response to background noise and N is the noise level measured inside the car in dB SPL (A). Therefore, in addition to the three test cases, the system evaluates the classification accuracy considering the Lombard effect.

The noise levels measured at the three test cases are given in Table 1. The highest speech level recorded is 78.6 dB SPL and the lowest speech level recorded is 68.2 dB SPL. The recorded speech levels cover a wide range of SNRs from 0 - 25 dB.

Table 1: Background noise level at different test cases

Test cases	Noise level (dB SPL A)
Idle	51.8
60 km/h	63.7
120 km/h	67.9

Speaker Identification System

Figure 2 represents the tasks involved in speaker identification. Initially the speech data is collected from different speakers. A pre-processor divides the data into 20 ms frames and passes through a window function. A simple voice activity detector selects only the speech signals and the features are extracted. 60% of the data is dimension reduced and considered to develop the classification model. The remaining 40% of the data is used to test the developed model by evaluating the prediction accuracy.

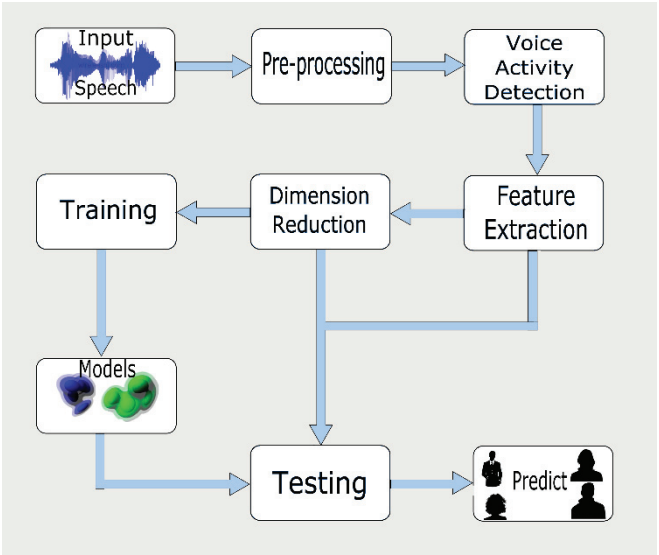

Figure 2: A representation of workflow in the speaker identification system.

Results analysis

In order to develop a system that is efficient in terms of computation and time consumption, it is essential to find optimal values for crucial parameters.

Time Duration: The classification accuracy is initially calculated for speech signals of 20 ms frame duration. Since 20 ms is a tiny frame duration to convey any information through speech, the accuracy is tested for different time duration using majority-voting. This algorithm selects the

most identified speaker within a time duration and assigns him or her as the main speaker for that time duration. Figure 3 shows a plot of average classification accuracy across all the speakers, obtained using majority voting, against different time durations of speech signals.

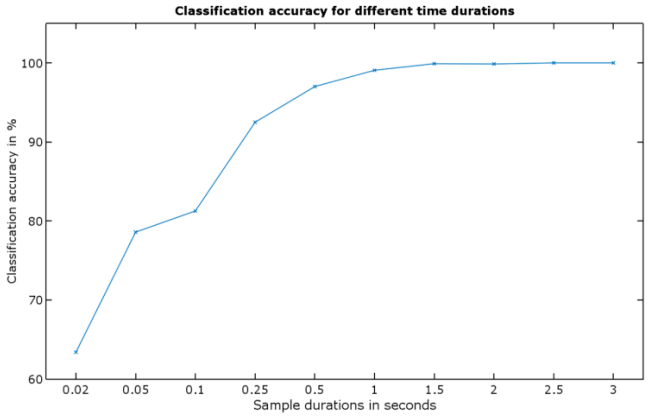

Figure 3: Classification accuracy for different time duration using majority-voting algorithm.

It is observed from the plot that the classification accuracy increases with the time duration and reaches a saturation point at 1 second. Therefore, 1 second is considered optimal for further calculations.

Gaussian components: As the GMMs are used, it is important to select the optimal number of Gaussian components that can be used for the task. Using fewer components leads to a case known as 'under-fitting', where the model is not good enough even to fit the training data. Similarly, using more components can result in 'over-fitting', where the model learns the training data in detail, which might lead to poor performance on the test data. Figure 4 shows a plot of average classification accuracy over all speakers (blue) and the minimum accuracy observed for the speaker with lowest identification accuracy (red) against different number of Gaussian components.

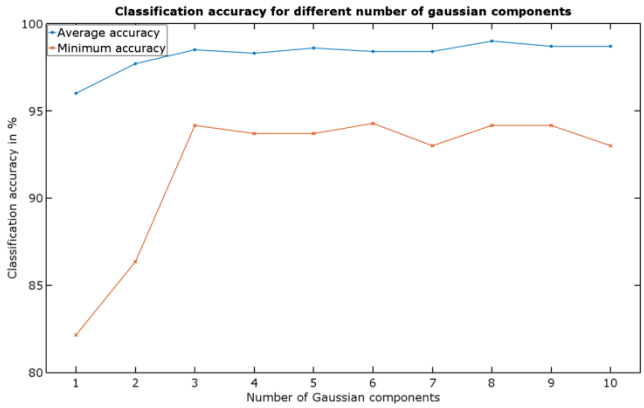

Figure 4: Average & minimum classification accuracy for different number of Gaussian components.

The average accuracy gradually increases for more components and reaches a saturation point and the minimum accuracy follows a similar behaviour. It is essential to build a model that not only gives a good average accuracy, but also provides a good identification accuracy for individual speakers. Therefore, an optimal of 6 Gaussian components are selected.

Feature-set dimensions: One of the factors that affects the computation time of the system is the dimension of the feature-set, as higher dimensions results in higher time consumption. Figure 5 shows a plot of average classification accuracy over all speakers (blue) and the minimum accuracy observed for the speaker with lowest identification accuracy (red) against different dimensional feature-set.

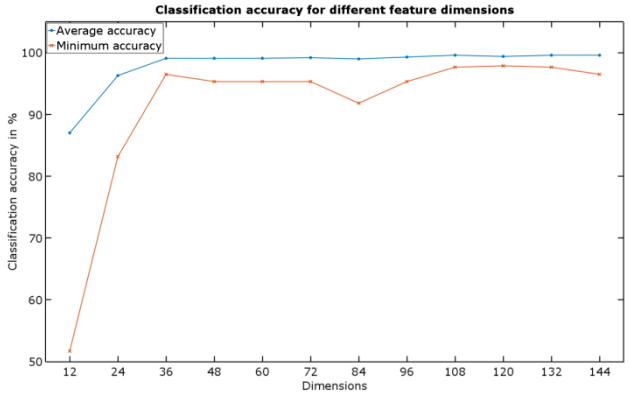

Figure 5: Average & minimum classification accuracy for different dimensional feature-set.

The average accuracy reaches a maximum for 36 dimensions and remains the same for higher dimensions. Even though the minimum accuracy is higher for higher dimensions, it is practically inefficient to use them for real-time implementation. Therefore, an optimal of 36 dimensions are selected.

Identification accuracy for test cases: By considering the defined parameters, the average accuracy's across all speakers obtained for all the test cases are represented in Table 2.

Table 2: Classification accuracy's for all test cases

Test cases	Average accuracy in % using majority voting
Clean speech (idle)	99.1
At 60 km/h	99.7
At 60 km/h with Lombard effect	99.8
At 120 km/h	97.7
At 120 km/h with Lombard effect	98.3

When the developed system is tested against clean speech signals, 99.1% of the time it identified the speakers correctly. Similarly, an accuracy of 99.7% and 97.7% is obtained when the car is moving at 60 km/h and 120 km/h respectively. Additionally, when the Lombard effect is considered in the presence of background noise, the accuracy increased to 99.8% and 98.3% respectively.

Conclusion

A real-time noise-robust speaker identification system is developed using GMMs as the classifier. 6 Gaussian components are opted for this task with a feature-set of 36 dimensions. A wide range of SNRs is considered from 0 - 25

dB and the speaker is predicted for every single second with an accuracy of 97.7%. Additionally, Lombard effect is found to increase the classification accuracy by 0.5 - 1%.

Some of the future works can be to assess how the classification accuracy varies with higher number of speakers, to identify "barge-ins" (sudden change in speaker) and to detect two or more simultaneous speakers.

Acknowledgement

Our sincere thanks to all the colleagues at IAV GmbH for their patience, help and support in collecting the speech data required for the task.

Literature

[1] D. A. Reynolds, “An Overview of Automatic Speaker Recognition Technology” in 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. IV-4072-IV-4075, May 2002.

[2] E. B. Tazi, A. Benabbou, and M. Harti, “Efficient text independent speaker identification based on GFCC and CMN methods,” in 2012 International Conference on Multimedia Computing and Systems, pp. 90–95, May 2012.

[3] T. Kinnunen, V. Hautamäki, and P. Fränti, “Fusion of spectral feature sets for accurate speaker identification,” in SPECOM’2004: 9th Conference "Speech and Computer", pp. 361–365, September 2004.

[4] R. S. S. Kumari and S. S. Nidhyananthan, “Frequency-time analysis approach to feature extraction for text independent speaker identification,” in 2011 International Conference on recent Advancements in Electrical, Electronics and Control engineering, pp. 258–262, December 2011.

[5] D. Schwarz and X. Rodet, “Spectral Envelope Estimation and Representation for Sound Analysis-Synthesis,” in 1999 International Computer Music Conference Proceedings, pp. 351-354, June 1999.

[6] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, “Linear discriminant analysis: A detailed tutorial,” AI Communications, vol. 30, no. 2, pp. 169–190, May 2017.

[7] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72–83, Jan 1995.

[8] H.-G. Kim, N. Moreau, and T.Sikora, “Mpeg-7 Audio and Beyond: Audio Content Indexing and Retrieval”, John Wiley & Sons, 2005.

[9] N. Hanai, and M. S Richard, “Robust speech recognition in the automobile,” Proceedings of the ICSLP, pp. 1339-1342, September 1994.

[10] E. Lombard, “Le signe de l’élévation de la voix [The sign of the rise in the voice],” Annales des Maladies de l’oreille, du Larynx du Nez et du Pharynx, vol. 37, pp. 101–119, February 1911.

[11] Recommendation ITU-T P.1110, Wideband hands-free communication in motor vehicles, March 2017.