On the Use of Artificial Conversation Data for Speaker Recognition in Cars

Luke R. Gottlieb
International Computer Science Institute
Berkeley, CA 94704 USA
luke@icsi.berkeley.edu

Gerald Friedland
International Computer Science Institute
Berkeley, CA 94704 USA
fractor@icsi.berkeley.edu

Abstract

Based on contrastive experiments the following article presents a discussion on the difficulties of using state-of-the-art speaker recognition methods under realistic car noise conditions and argues that, even though work has been done in this area, current approaches fail to address the main problems occurring in this task. The article also proposes a possible solution which opens new directions for further research in this area.

1. Introduction

Audio-based recognition technologies are increasingly used in everyday environments. Recently, speaker recognition and diarization in cars raised the attention of researchers because the car industry has become interested in command-language speech recognition in high-end cars. One particular problem is, that for safety reasons, certain commands should only be executed when uttered by the correct person in the car. For example, consider a family riding in a car: Children might be disallowed to open certain windows by command or adjust the current route of the GPS system. A parent, even when sitting in the back seat and taking care of a younger child, however, should be able to use the command recognizer to change the route of the GPS or open the windows. The enabling technology for such a user interface is researched in the field of speaker identification and recognition. This article shows that state-of-the-art speaker recognition systems are not suitable for speaker recognition in car-noise conditions. By performing contrastive experiments with both an artificially generated car-noise corpus and a real car-noise corpus, we provide empirical evidence on how much car noise conditions affect the performance of speaker recognition and that state-of-the-art approaches do not address these issues. Our experiments also point to a direction for future research in this area. Section 2 presents related work on speaker recognition and car-related speech and speaker recognition. The article then continues by discussing the fundamental properties of car noise in Section 3. Section 4 describes our data, and how it was collected and Section 5 details our speaker recognition system and the error metric which we used. Section 6 supports the hypotheses in this article by presenting quantitative results. Section 7 concludes the article and presents future work.

2. Related Work

The task of speaker recognition is usually distinguished in two categories. In the first one, a speaker claims to be of a certain identity and his or her voice is used to verify this claim. This is called speaker verification. In speaker identification, the task is to determine an unknown speaker's identity. One application of speaker identification includes looking at a criminal's voice and cross-checking it against a database of criminal's voices looking for a match. Speaker recognition systems employ three styles of spoken input: text-dependent, text-prompted, and text-independent. Most text-independent speaker identification systems use a GMM/UBM [16] approach, along with a variety of channel normalization techniques [2] (e.g., feature-, model-, and score-level). Speaker identification is regularly evaluated by NIST. However, given their distinct field of application, speaker identification systems are tuned to require several seconds of input data. In the classic speaker identification scenario, many utterances of speech from a large number of different people are given, and the task consists of mapping each utterance to each speaker. Common training sets consist of many hours of speech [5], and the test data must usually be several tens of seconds long. Five seconds is considered a very short utterance. Most of the techniques are therefore not suitable for the use with command language.

Even though there is current interest in speaker identification approaches in cars [19, 14], there seem to be no publicly available speech corpora in cars that is annotated for speaker identification. Car speech corpora seem to concentrate on single speaker speech recognition, the most popular one being [10]. As a result, speaker identification research

978-0-7695-3800-6/09 $26.00 © 2009 IEEE
DOI 10.1109/ICSC.2009.35

seems to almost exclusively report results on standard corpora "enriched" with car noise[1], examples are [7, 6, 17]. Mostly, the solution reported includes applying a set of self-designed filters, e.g. as presented by [13]. Based on the scenario presented in Section 1 this article will discuss the relevance of these approaches.

3. Properties of Car Noise

To explain the difficulties of speaker identification in cars this section will briefly explain the properties of the task. Conversations in cars and in meeting rooms differ in various ways. First, regular conversations, e.g. in a business meeting, are usually focused around one person that has the floor and the remaining people listening to the speaker. Even though overlap occurs, it is usually rare and depends on the cultural acceptance and social status of the meeting participants [3]. In a car it is culturally more acceptable for the persons in the back seat to have a separate conversation from the persons in the front. Therefore, speech overlap is a much larger issue. The topic of the conversation can easily be interrupted by the general topic of finding the way or by external events such as a misbehaving car. While this might not affect speaker identification performance very heavily, it definitely contributes to the differences between meeting recordings and car conversations. On the other hand, in car conversations, people are unlikely to change their location, in contrast to meeting recordings, where participants might stand up, walk around, change seats and so forth.

The noise in cars is very diverse. Most of the distortion comes from the running motor. Given a specific model, the volume and pitch of the motor sound is dependent upon the speed of the car. The motor sound is also modulated by the environment, e.g. a car driving through a tunnel has a rather different sound than the same car on an open street. The pavement of the street adds additional noise and adds to the modulation of the motor sound. The second most important distortion comes from in-car sounds: The turn signal ticks, the windows open, the toll system beeps, the windshield wipers scratch, and so on. Finally, environmental noises add to the scene: Rain drops on the car roof, vehicles honk, breaks squeal, people shout, etc.

The signal-to-noise ratio of these distortions is difficult to measure and ultimately depends upon the position of the microphone in the car, the car body, and the environment in which the car is driving. Therefore, adding artificial noise to meetings is definitely not an optimal simulation. Also, adding artificial noise to meetings bears the problem of not taking into account the McGurk [12] and Lombard effects [9]: Speech is impacted by surrounding noise and also visual effects in the environment. It can therefore become much faster, louder, and even change pitch relative to the happenings surrounding the speaker.

We therefore believe that it is very difficult in the short term to find a generic filter for this scenario and therefore propose a different solution, that is presented in the next section.

[1] A similar situation had been reported for speech recognition approaches before corpora existed there, see [11].

Figure 1. An iPhone mounted to the car windshield running the program "Recorder" was used to record the noise and car conversations, simulating a built-in microphone in the car.

4. Data Collection

In order to perform the experiments presented in this article, we used two main data sources.

4.1 Baseline: AMI corpus

As a baseline and contrastive condition, we used a selected set [8] of meetings from the AMI corpus [4]. The AMI corpus consists of audio-visual data comprised of four participants in a natural meeting scenario. The participants volunteered their time freely and were assigned roles such as "project manager" or "marketing director" for the task of designing a new remote control device. The teams met over several sessions of varying lengths (15-35 minutes). The meetings were not scripted and different activities were carried out such as presenting at a slide screen, explaining concepts on a whiteboard or discussion while sitting around

a table. The participants therefore interacted naturally, including talking over each other. The meeting recordings, however, were collected in an instrumented meeting room. In order to simulate a car conversation, we mixed car noise to the data. The car noise was recorded as described in Section 4.

4.2 Car Data Collection

As already described in Section 2, very few publicly available car noise corpora exist and certainly none that are annotated for speaker identification. In order to experiment with real data we recorded about 3 hours of car noise and 1 hour of annotated 4-person car conversation. We collected two sets of car recordings by mounting an iPhone to the windshield of an automatic Dodge Neon. We recorded the morning and evening commute from Berkeley to San Francisco and back. The recordings contain any noises occurring during the driving, including different motor noise levels from speed between 0 and 75 miles per hour, turn signal ticking, doors and window opening, and outside noise like horns and break squeals. The recordings do not include rain, windshield wiper noise, or radio programming. We used a program called "Recorder" which is available for the iPhone and records at a 44 kHz sampling rate, Figure 1 shows a photo of the setup. The car conversations were recorded using the same setting except that four people were freely engaged in a conversation in the car. The driver was asked to go to a place he had never been before and therefore had to ask for directions from time to time, including finding a spot for parking.

5 Online Speaker ID Approach

For the experiments reported herein, we use a variation of the approach presented in [18]. The approach is different from most speaker identification systems in that it can deal with very sparse input data: About 2-2.5 seconds of speech is sufficient to identify a speaker, compared to the usual 5-10 seconds (see Section 2). The purpose of the system is to answer the questions "is somebody speaking now?" and if so: "who is speaking?". For the system to work online, the questions have to be answered on small chunks of the recorded audio data, and the decisions must not take longer than realtime. Figure 2 shows a big-picture overview of the system. The steps are described as follows.

5.1 Training

In training mode, one minute of an audio sample is used for training. Either the speech of a single speaker is used or different noise conditions are trained that do not contain any speech. The audio is converted into 19-dimensional MFCC features. The segment is then used to train a Gaussian Mixture Model (GMM). The optimal number of Gaussians and iterations has been determined empirically as 20 Gaussians per model, as described in [18].

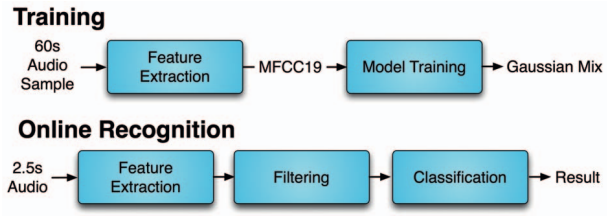

Figure 2. The main steps of the described system as outlined in Section 5.

5.2 Recognition

In the actual recognition mode, the system records and processes chunks of audio as follows. In a first step of feature extraction, the sampled audio data is converted into 19th-order MFCC features. Cepstral Mean Subtraction (CMS) is implemented to help deal with stationary channel effects. Although in subtracting the mean some speaker-dependent information is lost [15], according to the experiments performed, the major part of the discriminant information remains in the temporal varying signal. For some of the experiments, we also used a Wiener filtering approach [1] as an additional filtering step (see Section 6).

In the classification step, the likelihood for each set of features is computed against each set of Gaussian Mixtures obtained in the training step. As determined by the experiments on larger meeting corpora (see [18]), we use 2.5-second chunks of audio and a frame-length of 10 ms. This means, a total of 250 frames are examined to determine if an audio segment belongs to a certain speaker in the non-speech model. The decision is reached using majority vote on the likelihoods.

All the above steps are computed on the fly, requiring less than 10 % real time using a Macbook Pro.

5.3 Error Metrics

For evaluation, the system is run offline with different training and testing conditions. In this mode, the output of the system consists of metadata describing speech segments in terms of starting time, ending time, and speaker cluster name. This output is then evaluated against manually annotated ground truth segments (see Section 4). A dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the

System	DER
No car noise in training or recognition set	21.28 %
Car noise only in recognition set	49.81%
Car noise in training and recognition	37.39%

Table 1. A comparison of experiments performed on a subset of the AMI meeting data with car noise added at different stages.

System	DER
Noise only in recognition set	54.27%
Noise only in recognition set and filtering	66.74%
Noise in training and recognition	46.82%
Noise in training and recognition and filtering	42.03%

Table 2. A comparison of experiments performed on real car conversation data with car noise and filtering added at different stages.

ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate which is defined by NIST[2]. The Diarization Error Rate (DER) can be decomposed into three components: misses (speaker in reference, but not in hypothesis), false alarms (speaker in hypothesis, but not in reference), and speaker-errors (mapped reference is not the same as hypothesized speaker). The speaker error includes all wrongly classified segments, including overlapped speech and very short segments.

6. Experimental Results

6.1 Adding Artificial Car Noise to Meeting Recordings

This section presents the contrastive experiments performed using the data and system presented above. As a baseline, we ran the online speaker identification approach, as described in [18] on a subset of the AMI meetings without adding car noise. The diarization error rate for the single-distant microphone case and four speakers is 21.28 %, which includes a speech/non-speech error of 12.2 %. We used the same speech segments throughout the experiments in order to measure only the speaker identification performance. Therefore, all of the experiments reported in Table 1 contain the same speech/non-speech error of 12.2 %. Adding car noise only in the recognition step, increases the error to about 50 % DER. If speaker models are trained with mixing a random selection of car noise to the 60-second samples, the accuracy improves dramatically and error drops down to 37.39 %.

6.2 Real Car Experiments

As described in Section 3 there are many reasons why adding car noise artificially does not reflect the conditions of the car situation properly. Therefore, we repeated the experiments with the real conversations recorded in a car (as presented in Section 4). First, when comparing Table 1 and 2, one can see that performance on real car data is worse overall than the performance on artificial data. Both the artificial data and the real data contain the same number of speakers (4) and are about the same length. Again, when training with clean speaker samples, the DER is much higher than when training with randomly selected car noise mixed into the speaker set. This remains true even if one uses Wiener filtering [1]. The filtering only improves the results when the the speaker models were trained with real car-noise mixed into the speaker samples.

7. Conclusion and Future Work

This article presented a basic discussion of the properties of car conversation data in relation to speaker recognition. Using an off-the-shelf online speaker recognition system, empirical evidence is provided that shows both the difficulties of the task as well as a potential direction for a solution. Given the observations made in Section 3 as well as the quantitative results of Section 6, we suggest to base any investigation of car-based speech and speaker recognition system exclusively on real-world car recordings. In addition, we think that systems will benefit markedly from pretrained noise models that might be built specifically for each car model and type. The discussion in this paper provides evidence that current work has not yet taken the problems in car environments very seriously and/or seems to go in a not so promising research direction as related work almost exclusively discusses the use of generic noise filters (most of them stationary) on artificial data. We suggest that car conversations be recorded and annotated with different models of cars driving under different environmental conditions. The corpus should be open to the research community for experiments to be compared freely. Only then will we begin to understand the effects and impacts of complex noise environments on speech data, and in the end, contribute to the successful semantic analysis of the information contained therein.

[2]http://nist.gov/speech/tests/rt/rt2004/fall

References

[1] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas. Qualcomm-icsi-ogi features for asr. In *Seventh International Conference on Spoken Language Processing*. ISCA, 2002.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacraz, and D. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.

[3] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4353–4356, 2008.

[4] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraiij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, M. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus: A pre-announcement. In *Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, 2005.

[5] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE ICASSP*, 1992.

[6] S. Goetze, K.-D. Kammeyer, and V. Mildner. Multi-channel noise-reduction-systems for speaker identification in an automotive acoustic environment. In *Proceedings of the Audio Engineering Society Convention*, page 6756, May 2006.

[7] H. Guangrui and W. Xiaodong. Improved robust speaker identification in noise using auditory properties. In *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*, pages 17–19, 2001.

[8] H. Hung, Y. Huang, C. Yeo, and D. Gatica-Perez. Correlating audio-visual cues in a dominance estimation framework. In *CVPR Workshop on Human Communicative Behavior Analysis*, 2008.

[9] J. Junqua, S. Fincke, and K. Field. The Lombard effect: a reflex to better communicate with others innoise. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings.*, volume 4, 1999.

[10] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagaki. Construction of speech corpus in moving car environment. In *Sixth International Conference on Spoken Language Processing*. ISCA, 2000.

[11] I. Lecomte, M. Lever, J. Boudy, and A. Tassy. Car noise processing for speech input. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 512–515 vol.1, May 1989.

[12] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 1976.

[13] J. Meyer and K. Simmer. Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1167–1170 vol.2, Apr 1997.

[14] C. Mueller and G. Friedland. Multimodal interfaces for automotive applications (miaa). In *Proceedings of ACM IUI 2009*, pages 493–494, February 2009.

[15] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108, 1995.

[16] D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proc. of International Conference on Audio and Speech Signal Processing*, 2005.

[17] E. Vale, A. Cunha, and A. Alcaim. Robust text-independent speaker identification using multiple subband-classifiers in colored noise environment. In *Systems, Signals and Image Processing, 2008. IWSSIP 2008. 15th International Conference on*, pages 275–278, June 2008.

[18] O. Vinyals and G. Friedland. Towards semantic analysis of conversations: A system for the live identification of speakers in meetings. In *Proceedings of the 2nd IEEE International Conference on Semantic Computing*, pages 426–431, August 2008.

[19] W. Wahlster. Smartweb: multimodal web services on the road. In *ACM Multimedia*, page 16, 2007.