ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ADDITIVE AND CONVOLUTIONAL NOISES COMPENSATION FOR SPEAKER RECOGNITION

Olivier Bellot, Driss Matrouf, Teva Merlin and Jean-François Bonastre

Laboratoire Informatique d'Avignon (LIA)
Agroparc BP 1228,
84911 AVIGNON Cedex 9, FRANCE
{olivier.bellot, driss.matrouf, teva.merlin, jean-francois.bonastre}@lia.univ-avignon.fr

ABSTRACT

It is well known that the performances of speaker identification systems degrade rapidly as the mismatch between training and test conditions increases. In this work we present a noise compensation technique whose goal is to minimize the effects of such mismatch, so as to obtain an identification accuracy as close as possible to that obtained under matched conditions. To reduce this mismatch, the adopted approach compensates the speaker model parameters using the noise present in the test data, and compensates the test data frames using the noise present in the training data. The test and the training data (for different speakers) are assumed to come from different and unknown microphones and acoustic environments.

1 INTRODUCTION

The performance of an automatic speaker recognizer degrades drastically when the recognizer is used in the environments mismatched to the training environments. The goal of noise compensation is to minimize the effects of such mismatch, so as to obtain a recognition accuracy as close as possible to that obtained under matched conditions. In real world signal acquisition systems, the acquired signal is often a distorted version of the source signal that has been contaminated by both additive and convolutive noises. In speaker recognition applications, the mismatch between training and test data might be attributed to telephone network transmission or recording using different microphones. In dealing with additive noise, the approaches range from speech waveform enhancement to acoustic model adaptation [1,2,3,4,5]. To deal with convolutive noise, recent approaches have focused on the removal of cepstral bias [6,7]. To deal with the simultaneous presence of additive and convolutive noises, successful approaches include the SNR dependent compensation of acoustic models [8], the approximation method in the log spectral domain [9], and an iterative method for bias removal in both the spectral and cepstral domains [10]. In this paper, we describe a technique which compensates additive and convolutive noises using an iterative framework. The additive noise is compensated in the spectral domain and the convolutive one in the cepstral domain. This technique has proved to be effective in a speech recognition frame-work [11].

Approaches based on channel characteristic modeling generally assume that the training data is noise-free and the test data is noisy. In practice, this assumption is rarely correct. In this paper, we investigate an iterative procedure to compensate noises both in training and test data. Training and test noises are expected to be of different natures with different levels. Data are also assumed not to be recorded under the same conditions and are likely to come from different and unknown acoustic environments.

The techniques described in this paper have been assessed using the LIA speaker recognizer, AMIRAL [14], based on Gaussian Mixture Models (GMM). The experiments were carried out on BREF corpus - clean speech with a artificially added noise - and on Switchboard - multi-session telephone records of conversational speech.

In the next section, we present the proposed compensation method. Two situations are described: first when only test data are noisy and the latter when both training and test data are noisy. Section 3 presents the experiments using both a clean database (and artificially added noise) and a telephone multi-session database. The last section (4) is dedicated to some conclusions and comments.

2 Channel and noise compensation process

2.1 Clean training data, noisy test data

In this section we assume that only test data are noisy; i.e. there is no additive noise in training data. We also assume that the mismatch attributed to the channel can be represented by linear filtering.

In this case, the training and test signals can be modeled as follows:

(eq. 1) for training data $y_1 = h_1 * s$
(eq. 2) for test data $y_2 = h_2 * (s + n_2)$

where *s* is the hypothetical noise-free signal, h_1 corresponds to the channel effects in the training data (convolutional noise), n_2 and h_2 represent respectively the additive noise and the channel (convolutional) noise in test data.

To reduce the mismatch between training and test data, our approach compensates the speaker model parameters using the noise present in the test data. This can be done using a parallel model combination (PMC) technique [5] which approximates noisy speaker models by combining a speaker model (trained on y_1) with the noise model (trained on n_2). Various PMC techniques have been proposed: log-normal approximation [5], numerical integration [12] and data driven approaches [13,11]. In this work, we have chosen a data driven approach [11]. The approximations required for model combination are avoided by using the original training data instead of using only the associated models or generating speech frames from speaker models.

Using an appropriate operating scheme, the combination processes based on original training data can be performed as fast as PMC techniques. For a speaker model, the combination process is carried out sentence by sentence, using the following sequence of operation:

- The GMM speaker model is trained using the original training data. Any estimation algorithm can be used for this task.
- The association between each training frame and the corresponding (speaker model) Gaussian components is determined and stored once - before the compensation process - and remains unchanged during the combination process. We have chosen to associate each training frame with the Gaussian component corresponding to the highest posterior probability.
- The compensation process is now applied. This process is described in the next paragraph.
- Lastly, the stored association (memorized during the second phase) is used to obtain the noise compensated models. For each Gaussian component, a new mean vector and a new covariance matrix are directly calculated.

For the noise compensation only frames representing h_2*n_2 are available according to the channel model described by equations (1,2). To combine the training data utterances with the test noise frames, we need an estimate of frames representing h_1*n_2. To obtain h_1*n_2 frames, we estimate filter $h_1*h_2^{-1}$ which is applied to h_2*n_2 frames following equation 3:

(eq. 3) $(h_1*h_2^{-1})*h_2*n_2 = h_1*n_2$.

Filter $h_1*h_2^{-1}$ is obtained using an iterative process, in the cepstral domain. Equation 4 shows the form of this filter:

(eq. 4) $h_1*h_2^{-1} = \text{mean}[(s+n_2)*h_2] - \text{mean}[(s+n_2)*h_1]$

The first estimate of n_2*h_1 is h_2*n_2. The normalization is achieved by subtracting the cepstral mean from the modified training data (after combination) and from the test data frames. This normalization respects the convolutional component.

Three remarks complete the description of the proposed method:

- The combination of training data with the test noise is done in the spectral domain - frame by frame - in a circular manner. This technique allows proper and straightforward adaptation of covariance matrices.
- Theoretically, if there is no additive noise in test data (n_2=0), our compensation process is equivalent to the cepstral mean removal technique. Furthermore, if there is no channel mismatch, the algorithm must converge in one iteration.
- It has been shown that the use of delta coefficients improves speaker recognition performance. The compensation of noise for this kind of coefficient is difficult to perform using classical techniques. The proposed approach keeps the temporal context of a frame, since the original training data are used, which enables a very straightforward calculation of the noise compensated delta parameters.

2.2 Noisy training and test data

In real-world applications the assumption that training data are clean is rarely correct. If we assume that both training data and test data are noisy, then the channel model becomes :

(eq. 5) for training data $y_1 = (s + n_1) * h_1$
(eq. 6) for test data $y_2 = (s + n_2) * h_2$

In this case the compensation consists in combining the test noise with training data frames y_1 and the training noise with test data frames y_2. To do so, we need an estimate of h_1*n_2 (modified test noise) and an estimate of h_2*n_1 (modified training noise). To obtain these, we estimate filter $h_1*h_2^{-1}$ which is applied to h_2*n_2 frames to get an estimate of h_1*n_2. Using the same process, the estimated filter $h_2*h_1^{-1}$ is applied to training noise frames h_1*n_1 to obtain an estimate of h_2*n_1. The process is iterated until the filters converge. After combination training and test data are described as :

(eq. 7) for training data $y_1 = (s + n_1 + n2) * h_1$
(eq. 8) for test data $y_2 = (s + n_2 + n_1) * h_2$

Normalization - which respects the convolutional components - is done by subtracting the cepstral mean from the modified training and test data.

To create the modified cepstral test data, the mean of training noise frames is used.

3 EXPERIMENTAL RESULTS

Two series of experiments are conducted. The first experiments use clean speech with an artificial adding of noise in order to assess the method under controlled conditions. The latter experiment simulates a real telephony application.

In both cases, a close set speaker identification scheme is used in order to evaluate the different solutions. The evaluation criterion is the percentage of good identification (number of good test / total number of test).

We used the LIA reference system: AMIRAL. The speaker models are GMM with 128 Gaussian components, each represented by a mean vector, a diagonal covariance matrix and a weight. The models are estimated by EM algorithm, using an ML framework. The speech signal is represented by a 16 cepstral coefficient vector each 10 ms. No delta or delta-delta are used.

3.1 Data, noise protocol

Two different corpora are used:

- The first one (BREF) is a subset of Bref (read speech, recorded in a quiet environment). This subset contains records of 40 speakers. The speaker models are learned using two sentences totaling 30 s of speech. 120 tests (3 tests per speaker) are available, each lasting about 15 seconds. Data were recorded at 16 kHz.
- The second corpus (SWITCHB) comes from the NIST evaluation. It is a subset of the Switchboard database. The speech segments correspond to real telephone conversations, recorded during several sessions with different lines and environmental conditions.
  The experiments were conducted using 100 speakers. Each model was learned using 2 minutes of speech recorded during 2 different sessions. 190 tests were available with a mean duration of about 30s (per test).
  The records were digitized using an 8 kHz frequency.

The noise was added (for BREF experiments) by a direct addition of the noise signal to the corpora signals. Different signal/noise ratios (clean = 35 dB, 15 dB, 8 dB) are proposed. Two different noises are used: for training data, a white noise is used and for test data a real cocktail party noise is used.

3.2 Experiments using clean training and noisy tests (BREF)

Table 1 shows the results obtained with clean training data and noisy test data. Three comments can be made:

- The performance decreases drastically if no compensation is used (from 98.3 % of good identification using clean test data to 7.5 % at 8 dB).
- Using a noise compensation technique allows an important gain. The best technique (compensation of the additive noise only) obtains 95 % of good identification using noisy test data (8 dB) compared to 7.5 % without normalization. The two techniques proposed in this paper perform better than a classical cepstral mean subtraction (CMS) technique (95 % compared to 21.6 %).
- Compensation of both additive and convolutive noises does not seem to allow a gain compared to additive noise compensation. This point could be explained by the nature of the database. All the records were made using the same microphone and the convolutive component was not very different from one record to another.

	Test data SNR (Training at 35 dB)		
	35 dB	15 dB	8 dB
No compensation	98.3 %	38.3 %	7.5 %
CMS	98.3 %	70.8 %	21.6 %
Compensation of additive noise	98.3 %	97.5 %	95 %
Compensation of add. and conv. noises	97.5 %	95 %	88.3 %

Table 1: Identification ratio with “cocktail party” noise in test at different levels of SNR depending on different compensation techniques. Training data are clean. Close-set identification tests using BREF corpora, 40 speakers and 120 tests.

3.3 Experiments using noisy training and test data (BREF)

Table 2 shows the results obtained with both noisy test data and noisy training. Without noise compensation, the performance decreases when the test SNR decreases. Compensating only the test noise allows a limited gain (30.7% for 16.6%, using 8 dB SNR for test data) when the test data are noisy. If the data test are clean, the compensation of only the test noise does not bring any gain. Using the compensation of both training and test noises allows a large gain in terms of performance in all the situations. The performances are always comparable to those obtained using clean test and training data (91.6 % using 8 dB test and training data to 98 % using clean data).

	Test data SNR (training at 8 dB)		
	35 dB	15 dB	8 dB
No compensation	40.8 %	30 %	16.6 %
Compensation (test only)	40 %	35.8 %	30.7 %
Compensation (train and test)	97.5 %	94.2%	91.6 %

Table 2: Identification ratio with “cocktail party” noise in test at different levels of RSB depending on different compensation techniques. Training data is noisy (white noise and SNR=8 dB). Close-set identification tests using BREF corpora, 40 speakers and 120 tests.

3.4 Experiments using real telephony data (SWITCHB)

Using SWITCHB data (table 3), performance increases when a compensation technique is used (36.8% to 62 % using test noise compensation or 68% using a CMS technique).

No compensation	36.8 %
CMS	68.4 %
Compensation on test data	62 % %

Table 3: Identification ratio. Close-set identification tests using SWITCHB corpora, 100 speakers and 190 tests.

The noise compensation technique proposed here performs worse than a classical CMS technique. Two comments can be made to explain this result:

- The noise compensation was only applied on test data, for calculation time reasons. Switchboard data are recorded using numerous telephone lines and environmental conditions. It could be supposed that compensating both training and test data should increase the performances as seen in 3.3 (a 30 % to 91% increase is allowed by using both noise compensations).
- The noisy frame detection algorithm used here is very simple: we chose the 20% least energized frames. Finding a better algorithm is not easy but should allows a gain in performance.

4 CONCLUSION

To deal with the mismatch between training and test environmental conditions in speaker recognition applications, we have proposed a noise compensation method. This iterative method compensates additive and convolutive noises directly at the data level. The main advantages of this method are to allow the compensation of the noise present in both test and learning data, to take into account the variance of the different noises and to facilitate the use of delta coefficients.

A first set of experiments, carried out using a "clean" database (Bref) and a noise adding protocol, shows that noise compensation allows a large gain in terms of performance, particularly if both noise and test data are noisy (91.6% of good identification rate with 8 dB data, compared to 16.6% without normalization). In addition, the results obtained using clean speech segments are not disturbed by the compensation technique.

Using a telephone database (Switchboard/NIST), the results are not so satisfactory, compared to those obtained with a classical CMS technique. This problem seems linked to the noisy frame selection algorithm based on the energy level (low energy frames are considered as noise frames).

Future work will focus on the noise detection algorithm and on the adaptation of the method for speaker verification applications, particularly for NIST evaluations.

5. REFERENCES

[1] Y. Ephraim and H. L. Van Trees, "A signal Subspace Approach for Speech Enhancement", IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 4, pp. 251-266, July, 1995.

[2] S.F. Boll, "Suppression of Acoustic noise in Speech Using Spectral Substraction," IEEE Trans. on Acoustics Speech and Signal Processing, Vol. 27, pp. 113-120, April, 1979.

[3] A. Varga and R. K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," Proc. ICASSP, pp. 845-848, Albuquerque, NM, April, 1990.

[4] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated Models of Signal and Backgound with application to Speeker Identification in Noise," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 245-258, April, 1994.

[5] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 5, pp. 352-359, Sept.,1996.

[6] Y. Zhao, "An Acoustic-Phonetic-Based Speaker Adaptation Technique For Improving Speaker-Independent Continuous Speech Recognition," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 3, pp. 380-394, July, 1994.

[7] Y. Zhao, "Self-Learning Speaker/Channel Adaptation Based on Spectral Variation Source Decomposition," Speech Communication, Vol. 18, No. 1, pp. 65-78, Jan., 1996.

[8] A. Acero, R. M. Stern,"Environmental Robustness in Automatic Speech Recognition"' ICASSP, pp. 849-852, 1990.

[9] P. J. Moreno , B. Raj and R. M. Stern, " A Vector Taylor Series Approach for Environment-Independent Speech Recognition," Proc. ICASSP, pp. 839-842, Munich, Germany, April, 1997.

[10] M. G. Rahim and B.-J. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, pp. 19-30, Jan., 1996.

[11] D. Matrouf, J.L. Gauvain, "Model Compensation for Noises in Test and Training Data," Proc. ICASSP, pp. 831-834, Munich, Germany, April, 1997.

[12] M.J.F. Gales, S. Young, "Robust Speech Recognition in Additie and Convolutional Noise Using Parallel Model Combination," Computer & Speech language, 9(4), pp. 289-307, 1995.

[13] M.J.F. Gales, S. Young, "A fast and flexible implementation of Parallel Model Combination," Proc. ICASSP, pp. 133-136, 1995.

[14] C. Fredouille, J.-F. Bonastre, T. Merlin, AMIRAL: a block-segmental multi-recognizer approach for automatic speaker recognition, Digital Signal Processing, January 2000, Volume 10, 1-3