24th INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
23–27 July 2017, London

ICSV24
23-27JULY 2017
LONDON CALLING

COMPENSATING ACOUSTIC MISMATCH FOR ROBUST SPEAKER VERICATION

Víctor Poblete Ramírez
Isaac González Paulsen
Alexandra Astudillo Montenegro
email: vpoblete@uach.cl
Universidad Austral de Chile, Institute of Acoustics, P.O. Box 5111187, Valdivia, Chile

Gastón Vergara Díaz
Universidad Austral de Chile, Institute of Statistics, Casilla 567, Valdivia, Chile

Automatic speaker verification works better when the user speaks near the microphone in a noisy environment. Interaction with such systems may involve variations of speaker-microphone distance, a factor that together with additive noise of a room can dramatically decrease speech intelligibility and speech quality of recorded signal, causing a dramatic increase in the equal error rates (EERs). In this work, we extracted two sets of features: MFCC (Mel Frequency Cepstral Coefficients) and LNCC (Locally Normalized Cepstral Coefficients) to address the acoustic mismatch problem between training and verification environments. To analyze the robustness of these features to compensate for acoustic mismatches, several experiments of text-independent speaker verification (TI-SV) are performed with signals corrupted by additive noise at different signal to noise ratios (SNRs) along with different distances between loudspeaker and microphone within a same room. The reverberation time (T60) of an anechoic chamber is determined for four positions of loudspeaker-microphone distance. At each distance, versions of the YOHO speech corpus are re-recorded sequentially with a single microphone. Five types of noise are selected and recorded in the anechoic chamber. These noises are added to the YOHO versions to generate noisy signals of the utterances at various SNRs: 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB. We processed 3920 testing utterances x 4 distances x 5 noise x 6 SNRs = 470,400 signals. Our results indicate that LNCC provides relative reductions in EER, over standard MFCC. The highest reductions in EER are obtained with Airplane noise at SNR=10dB at a loudspeaker-microphone distance of 0.94 m, as high as 68% and 56% when compared with MFCC+CMN, or MFCC+RASTA processing, respectively.

Keywords: speaker verification, feature extraction, anechoic chamber, additive noise, distant speech.

1. Introduction

Speaker verification (SV) is a topic within the field of speech signal processing and refers to the automatic process through which we determine whether a given speech signal belongs to a claimed person or not (accept or reject, *i.e.*, a binary classification problem) based only on a voice sample [1]. SV has been used for remote security process for telephone banking, biometric security procedures, commercial purposes and many more. Automatic speaker verification (ASV) includes various steps: 1) acoustic feature extraction; 2) feature normalization; 3) speaker modeling (performed from the

extracted features); 4) model compensation; 5) verification; 6) score normalization; and 7) decision making. Two stages, enrollment and verification, constitute the SV system. In the enrollment stage, the training speech samples are processed by the feature extraction and speaker modeling stages to generate the match scores. In the verification stage, a testing utterance of a target speaker is also processed by the feature extraction stage, and then is matched with the trained models to decide whether it is a genuine speaker or not. Feature extraction aims to transform each speech signal to a set of feature vectors which provide enough discriminative information from the acoustic signal to enable the speaker to be verified, as plotted in Fig. 1 Feature extraction is the crucial input for statistical models. It is argued that a successful front-end feature extraction algorithm in ASV systems, should fit well the back-end speaker modeling, and be robust- by which we mean invariant- to changes of acoustic conditions. The accuracy of the ASV system is strongly dependent on this stage.

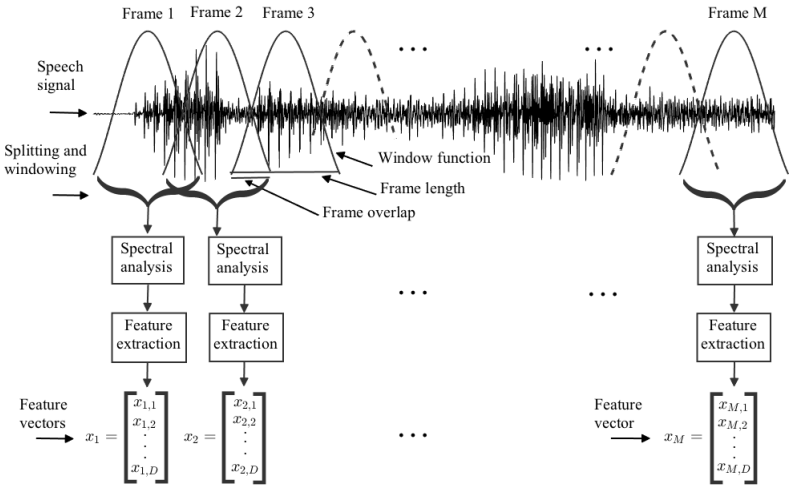

Figure 1: *The speech signal is segmented in frames using short and overlapped window functions.*

After feature extraction stage, feature normalization step is applied to filter the distortions, which is contaminated the extracted features. The statistical properties of the feature vectors could differ under the influence of noisy environments. The level of variation depends on the type of noise and also the level of contamination. Most of the normalization techniques are applied in the cepstral domain. Two of the most effective feature normalization schemes are Cepstral Mean Normalization (CMN) [1] and RelAtiveSpecTrAl (RASTA) [2]. CMN and RASTA are standard linear filtering compensation applied to the feature vectors prior to model training. Channel normalization is used to reduce environmental effects on the verification decision. Additionally, due to large distortion in ASV system, speaker model should also be adjusted by model compensation stage. This stage improves the SV performance by updating the speaker's model parameters in a more discriminative way. Research over recent years has demonstrated that Joint Factor Analysis (JFA) [3], one of the model compensation schemes, is capable to compensate for intersession variability and for channel mismatches between enrollment and verification conditions in particular. The extracted features from a speech signal are used as input to create a speaker model. We consider Gaussian Mixture Models (GMM) for speaker modeling. GMM has been applied extensively in SV systems. GMMs are probabilistic models, which assume that the acoustic feature vectors follow a Gaussian distribution. Under this approach, a speaker model, is created by an adaptation procedure from a Universal Background Model (UBM) or impostor model. GMM-UBM is a standard reference classifier in speaker verification [1]. Within this framework, expectation-maximization (EM), which is based on the concepts of maximum likelihood (ML), is typically used for model parameter estimation of UBM. Fig. 2 describes the two steps for training the GMM-UBM system. As can be seen from the Figure, the UBM is normally constructed from non-target speakers. After the construction of the UBM model, for each target model, a specific GMM is estimated by adapting the UBM *via* the Maximum a Posteriori (MAP) criterion (Fig. 2).

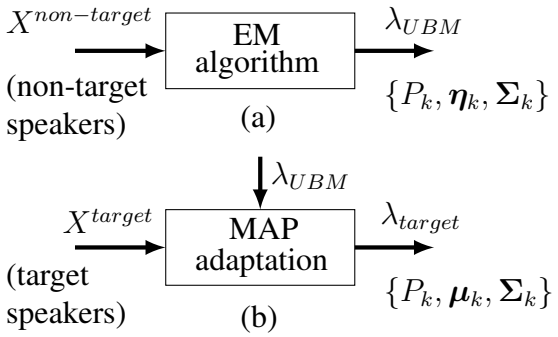

Figure 2: Steps of training a SV system. The GMM-UBM SV system is trained with (a)-(b).

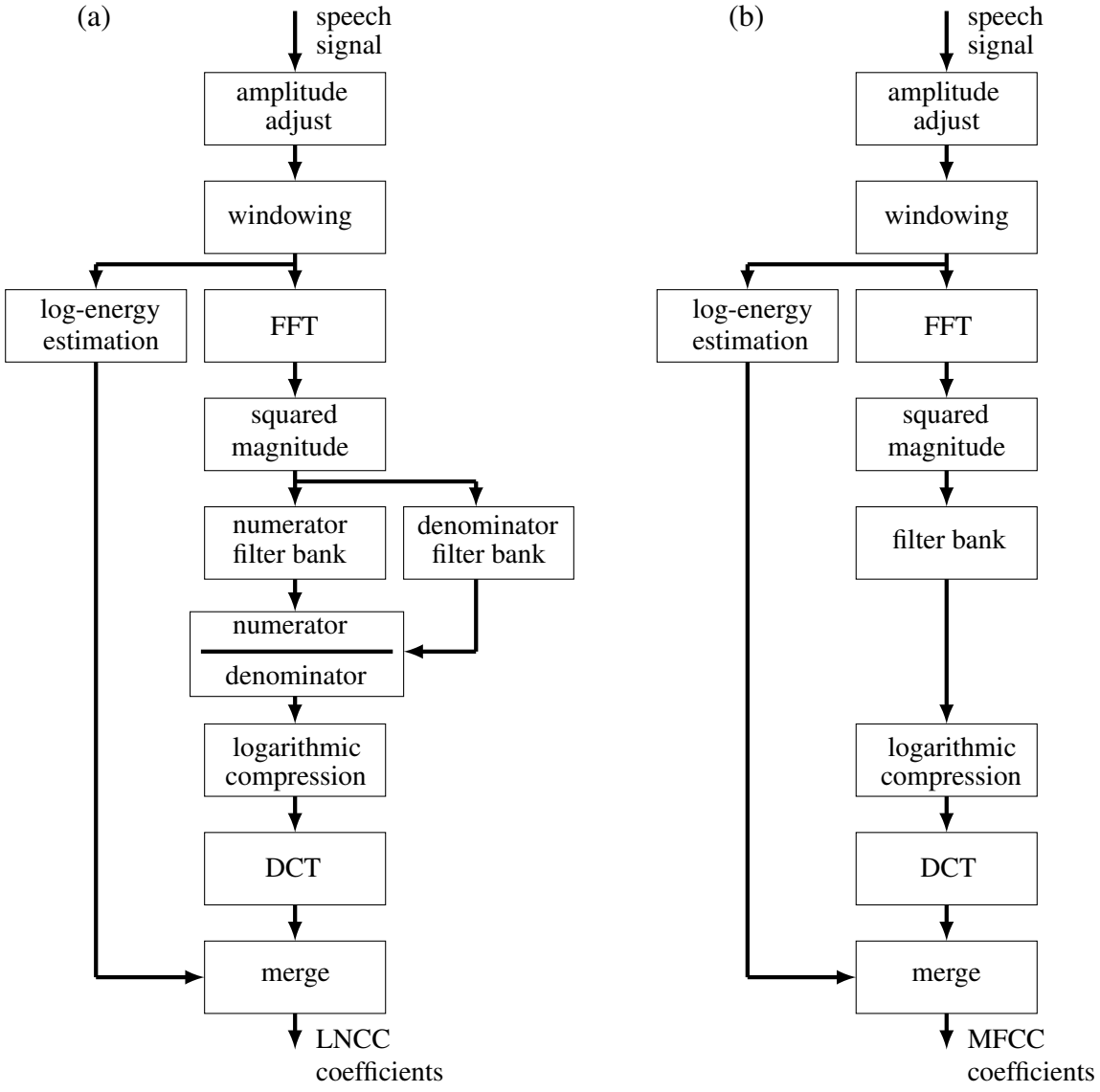

Figure 3: Diagram for: a) LNCC and b) MFCC feature extraction.

In the verification mode, the feature vectors extracted from an unknown speaker are compared against the models in the system database to give a similarity score. The term score refers to the log likelihood. The match score depends on both the target model (λ_{target}) and the world model (λ_{UBM}), as can be observed in Fig. 2. A central step in SV is how to make a decision. Essentially, an

ASV system could make two types of mistakes during decision-making: false acceptance (FA) that causes an impostor to be accepted, and false rejection (FR) which causes a genuine speaker's identity claim to be rejected. A substantial task in decision-making is somehow to minimize both FA and FR errors during decision-making. In the decision making stage, the score is compared to a threshold to distinguish between claimant and impostors speakers in verification. The performance of an ASV system can be evaluated using some metrics such as decision error trade-off (DET), equal error rate (EER), false acceptance rate (FAR), and false rejection rate (FRR). In this paper, we extract two sets of features: Mel Frequency Cepstral Coefficients (MFCC) and Locally Normalized Cepstral Coefficients (LNCC) to address the acoustic mismatch problem between training and verification environments. Then, we analyze the robustness of these features to compensate for acoustic mismatches. We will also perform several experiments of text-independent speaker verification (TI-SV) using signals from a re-recorded speech corpus, sequentially with a single microphone, in a real anechoic chamber, corrupted by additive noise at different signal to noise ratios (SNRs) along with different distances between loudspeaker and microphone within the same chamber.

The use of LNCC is based on Seneff's Generalized Synchrony Detector (GSD) [4]. The numerator is a triangular band pass filter centered around a particular frequency similar to the ordinary Mel filters. The denominator term is a filter that responds maximally to frequency components on either side of the numerator filter (see Fig. 3. It is common to append delta and delta-delta coefficients but this is not shown in the flowcharts). As a result, a local normalization is performed without the spurious peaks of the original GSD [6]. In the frequency domain, this local normalization is achieved by dividing the outputs of two filters. The numerator filter is triangular, and essentially the same as that used to derive MFCC features. While the denominator filter captures energy at adjacent frequencies. They are defined by the equations (1) and (2):

$$\mathrm{Num}_i(f) = \begin{cases} -\frac{2}{B}|f - f_i^c| + 1, & \text{when } |f - f_i^c| \leq \frac{B}{2} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$\mathrm{Den}_i(f) = \begin{cases} \frac{2}{B}\left(1 - d_{\mathrm{min}}\right)|f - f_i^c| + d_{\mathrm{min}}, & \text{when } |f - f_i^c| \leq \frac{B}{2} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

For any auditory filter i, the locally normalized filter energy LN_i is achieved by dividing:

$$LN_i = \frac{\mathrm{Num\_Energy}_i}{\mathrm{Den\_Energy}_i} \tag{3}$$

where $\mathrm{Num\_Energy}_i$ and $\mathrm{Den\_Energy}_i$ represent the energy captured by the filters in Eqs. (1) and (2).

2. Experiments

To investigate the ability of the LNCC features to reduce the mismatch between the training conditions and testing, over a wide range of types of noise, we carried out a sequence of text-independent speaker verification experiments on speech signals that are degraded by additive noise at different SNRs along with different distances between speaker and microphone. The experiments were carried out using the YOHO database [5]. We created several additive noise environments inside an anechoic chamber with a chamber volume of about 42m^3. In anechoic chamber, we re-recorded versions of the YOHO speaker verification corpus. The recordings were made using four speaker-to-microphone distances, from 0.47 to 3.76 m. YOHO database supports the development, training, and testing of speaker verification systems with a vocabulary comprising two-digit numbers spoken continuously in sets of three (e.g., "27-87-52" pronounced as "twenty-seven eighty-seven fifty-two"). YOHO was divided into training and testing sections. All the experiments were carried out using 138 speakers (32 females and 106 males), four training sessions per speaker with 24 utterances per session, and ten testing sessions per speaker with four utterances per session. The speakers were divided

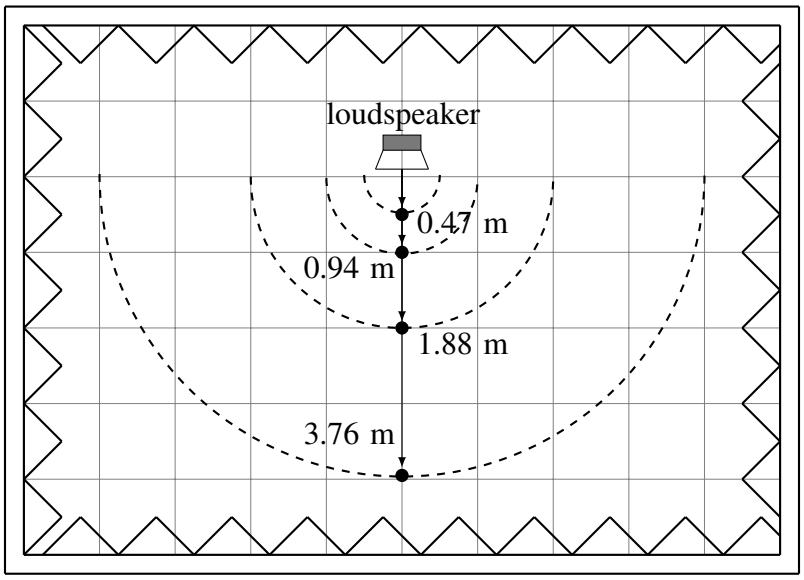

Figure 4: Re-recording positions. Black dot represents the position of the microphone.

as follows: 40 background impostor speakers to train the background models and 98 test client speakers for use in testing attempts. For each speaker, one 96-utterance training session was used. False rejection curves were estimated with 98 speakers$\times$40 testing signals per client = 3920 utterances. False acceptance curves were obtained with 98 speakers$\times$97 impostors$\times$40 testing speech signals per impostor = 380,240 experiments.

Four types of noise (airplane, restaurant, car, and mall) were selected from The Hollywood Edge Background Trax Sound Effects Library. Additionally, pink noise from STI-PA Signal Test, was also selected. These noises were reproduced by a loudspeaker (omni-direccional Cesva BP-012) in an anechoic chamber and re-recorded sequentially with a single microphone (Earthworks M30, omni-directional measurement microphone) at four distances (0.47, 0.94, 1.88 and 3.76 m) between the playback loudspeaker and microphone. During the recordings, Servo 260 power amplifier, AudioBox USB recording interface and a notebook HP 245 G4 model, were also used. These noise-versions were added to the YOHO corpus to generate noisy versions of the utterances at various SNRs: 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB. For all the speaker verification experiments, the system was trained with clean speech. The speech signal processed in total were 3920 testing utterances$\times$4 distances$\times$5 noise$\times$6 SNRs $=$ 470,400 speech signals. The results of the effectiveness of three conventional compensation techniques were also compared: CMN, RASTA, and JFA .

2.1 Text independent speaker verification system

The current paper basically relies on the information given by the GMM-UBM approach. The UBM represents the alternative hypothesis in the Bayesian test. It is designed to estimate the data probability not to belong to the targeted speaker. The UBM was learned with multiple speech signals from different background impostor speakers and was trained with the EM algorithm on its training data. For the speaker verification stage, UBM fulfills two roles: 1) it is the *a priori* model for all target speakers when applying Bayesian adaptation to obtain speaker dependent models; and 2) UBM helps to estimate log likelihood ratio by selecting the best Gaussian for each frame. Training was carried out using ALIZE-Open Source and LIA_SpkDet toolkits [7]. Speaker dependent models were derived by Bayesian adaptation on the Gaussian component means. The UBM was trained with 256 Gaussian components using diagonal covariance matrices. Features were extracted using LNCC and MFCC processing, as described by Fig. 3. The frame duration in all cases was 25 ms with a 50% overlap. A frequency range from 200 to 3860 Hz was covered by 14 triangular filters uniformly arranged on a

Bark scale, in the case of MFCCs, or 28 pairs of numerator and denominator filters uniformly arranged on a Bark scale in the case of the proposed LNCC features. If an LNCC filter goes beyond the range 0 Hz to Nyquist frequency, it is simply truncated. The DCT was truncated at 11 coefficients in both cases, then the first coefficient was replaced by the log frame energy. The resulting 11 coefficients are augmented with deltas and delta-delta coefficients to make up the final feature vector of dimension 33 for each frame. Differences between the LNCC and MFCC systems, including the compensation techniques, were evaluated by means of statistical hypothesis tests. Hypothesis tests used in the current paper evaluate the probability p that observed differences between groups of systems ocurred by chance, using McNemar's test. See for instance [8] for details about statistical hypothesis test and its application. The difference is considered as significant when the computed p-value is below a threshold, that is usually set to 0.001. The acoustic indices selected to quantify listening space in the anechoic chamber were reverberation time (T60), background SNR, and speech transmission index (STI), as listed in Table 1. Acoustic measurements were carried out with a sound level meter (Cesva SC310), the software Cesva Capture Studio and the acoustic analyzer Acoustilyzer AL1.

Distance (m)	SNR (dB)	T60 (s)	STI
0.47	41.3	0.09	0.94
0.94	38.7	0.09	0.94
1.88	31.7	0.11	0.85
3.76	22.0	0.10	0.74

Table 1: Acoustic indices.

3. Results

Fig. 5 describes results provided using average relative reductions in EER for speech in airplane noise and SNRs, using LNCC processing compared with the use of MFCC alone, and MFCC+CMN, RASTA, and JFA, at four distances. The experimental results with LNCC processing demonstrate a significant system performance improvement (p<0.001) over those with the conventional MFCC+CMN (Fig. 5). LNCC alone leads to average relative reductions in EER=68.3% and 67.1% (p<0.001), at SNR=10dB both for loudspeaker-microphone distances=0.47m and for 0.94m, respectively. It is worth noting that at the longest distance (3.76m), LNCC alone provides substantial and statistically significant relative reductions in EER as high as 54.4%, 50.2%, and 41.4% (p<0.001), compared to MFCC+RASTA, at 5dB, 10dB and 0dB, respectively. These results suggest that LNCC alone is far more robust than MFCC+RASTA, against changes in SNR for this type of noise at the largest loudspeaker-microphone distance. Despite this fact, we also note in Fig. 5 that, when LNCC processing is compared to MFCC+JFA, at the loudspeaker-microphone distance of 3.76m, at 20dB and 15dB SNR, LNCC exhibits poor performance in speaker verification.

Fig. 6 shows, for both processing: LNCC alone (left) and MFCC alone (right), the detection error tradeoff curves for the five types of noises at SNR=10dB, and at loudspeaker-microphone distance=3.76m, including clean speech conditions. In clean speech, at loudspeaker-microphone distance=1m, in the case of LNCC and MFCC, the EER are 1.1% and 0.71%, respectively. While for the five noise conditions, in the case of LNCC, at SNR=10dB, and at loudspeaker-microphone distance=3.76m, the EER are 13.2%, 18.1%, 16.7%, 13.04%, and 22.5%, for airplane, car, mall, restaurant, and pink, respectively. Now, in the case of MFCC, at SNR=10dB, and at loudspeaker-microphone distance=3.76m, the EER are 20.2%, 21.3%, 23.6%, 18.4%, and 38.0%, for airplane, car, mall, restaurant, and pink, respectively. Looking at these results, it appears clearly that the LNCC features compared to MFCC features, at the distance=3.76m plays an important role in the performance of SV. These results suggest that LNCC processing is far more robust than MFCC to distant speech. We now collate the results of the experiments, and consider the overall performance of LNCC alone compared to MFCC alone. Fig. 7 shows these collated results, using box-plots analysis of two

selected variables: loudspeaker-microphone distance and SNR, related to the EER results. We characterized the selected variable: loudspeaker-microphone distance (Fig. 7, left). As can be observed from the figure, the horizontal lines in the box interior, represent the median, and very little difference is observed, except for the loudspeaker-microphone distance, 3.76m, which shows in this case that LNCC features with median EER=19.6%, offer better performance than MFCC features with median EER=21.7%.

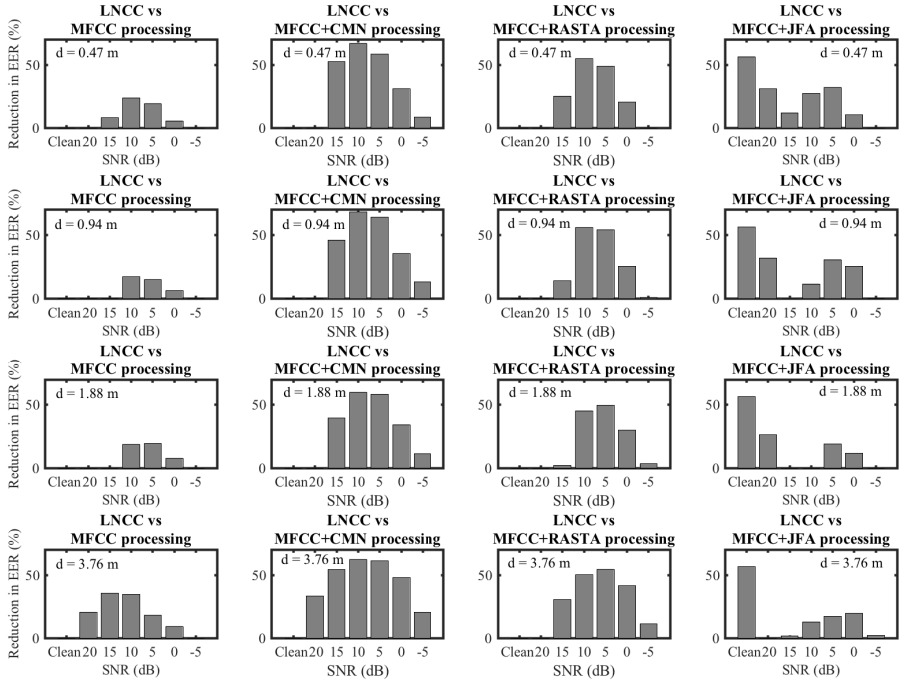

Figure 5: Relative reductions in EER for airplane noise.

By covering all loudspeaker-microphone distance conditions and the five types of noises, we characterized the selected variable: SNR (Fig. 7, right). As can be observed from the figure, the horizontal lines in the box interior, also represent the median, and again very little difference is observed, except at SNR=10 dB, which shows that LNCC features with median EER=16.8%, offer better performance than MFCC features with median EER=17.9%. We also observe the EER for clean speech condition only as reference points.

4. Conclusions

The robustness of the MFCC and LNCC features to compensate for acoustic mismatches have been compared for a speaker verification task across a wide variety of noisy conditions and with different distances loudspeaker-microphone. Our speaker-verification results demonstrate that: a) LNCC is more robust for largest loudspeaker-microphone distance that MFCC; b) CMN is not as helpful for LNCC; c) RASTA can improve the performance of LNCC; d) LNCC enables more robust speaker verification than MFCC+JFA approach. Additionally, LNCC is also, simpler and easier to implement than JFA; e) Our results indicate that at 10 dB SNR, covering all loudspeaker-microphone distance conditions and the five types of noises, LNCC in most cases, achieves statistically significant relative reductions in EER compared to the MFCC features. f) LNCC features can be an attractive alternative to MFCC, which can also be applied in other tasks of pattern recognition where occurs noisy environment, or when the loudspeaker-microphone distance is varying. LNCCs appear to be particularly attractive features for distant speech processing.

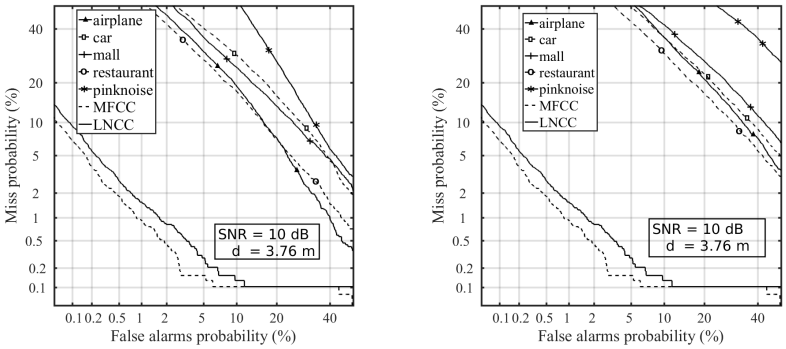

Figure 6: DET curves for LNCC features (left) and MFCC features (right).

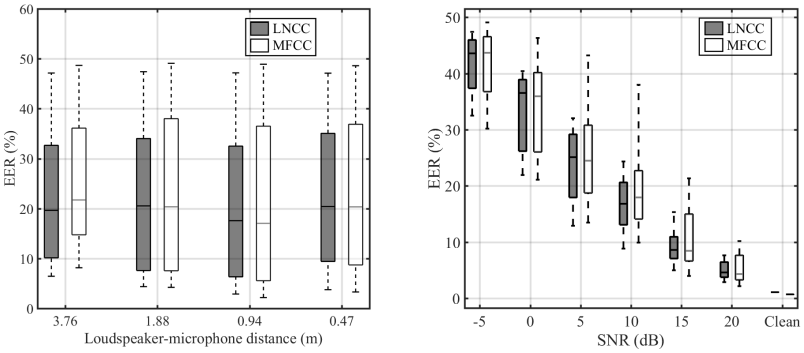

Figure 7: Overall performance for all tested conditions.

5. Acknowledgements

The research reported here was partly funded by grant DID-UACh 2015-63.

REFERENCES

1. Kinnunen, T. and Li, H., An overview of text-independent speaker recognition: From features to supervectors, *Speech Communication*, **52** (1), 12-40, (2010).

2. Hermansky, H. and Morgan, N., RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, **2**, 578-589, (1994).

3. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P., A study of interspeaker variability in speaker verification, *IEEE Transactions on Audio Speech and Language Processing*, **16**, 980-988, (2008).

4. Seneff, S., A joint synchrony/mean-rate model of auditory speech processing, *Journal of Phonetics*, **16** (1), 55-76, (1988).

5. Campbell, J. P., Jr., Testing with the YOHO CD-ROM voice verification corpus, *in International Conference on Acoustics, Speech, and Signal Processing ICASSP*, **1**, 341-344, (1995).

6. Poblete, V., Espic, F., King, S., Stern, R.M., Huenupan, F., Fredes, J., and Becerra, N., A perceptually-motivated low-complexity instantaneous linear channel normalization technique applied to speaker verification, *Computer Speech and Language*, **31**, 2-27, 2015.

7. Bonastre, J. F., Wils, F. and Meiner, S., ALIZE, a free toolkit for speaker recognition, *in International Conference on Acoustics, Speech, and Signal Processing ICASSP*, I-737–I-740, (2005).

8. Gillick, L. and Cox, S., Some statistical issues in the comparison of speech recognition algorithms, *in International Conference on Acoustics, Speech, and Signal Processing ICASSP*, **1**, 532-535, (1989).