Barlow Twins self-supervised learning for robust speaker recognition

Mohammad Mohammadamini, Driss Matrouf, Jean-François A Bonastre, Sandipana Dowerah, Romain Serizel, Denis Jouvet

Barlow Twins self-supervised learning for robust speaker recognition

Mohammad MohammadAmini[1], Driss Matrouf[1], Jean-François Bonatsre[1]
Sandipana Dowerah[2], Romain Serizel[2], Denis Jouvet[2]

[1]LIA (Laboratoire Informatique d'Avignon),
[2]University of Lorraine, CNRS, Inria, Loria, F-54000, Nancy, France

{mohammad.mohammadamini,driss.matrouf,jean-francois.bonastre}@univ-avignon.fr
{sandipana.dowerah,romain.serize}@loria.fr, denis.jouvet@inria.fr

Abstract

Acoustic noise is a big challenge for speaker recognition systems. The state-of-the-art speaker recognition systems are based on deep neural network speaker embeddings called x-vector extractor. A noise-robust x-vector extractor is highly demanded in speaker recognition systems. In this paper, we introduce Barlow Twins self-supervised loss function in the area of speaker recognition. Barlow Twins objective function tries to optimize two criteria: Firstly, it increases the similarity between two versions of the same signal (i.e. the clean and its augmented noisy version) to make the speaker embedding invariant to the acoustic noise. Secondly, it reduces the redundancy between dimensions of the x-vectors that improves the overall quality of speaker embeddings. In our research, Barlow Twins objective function is integrated with the ResNet-based speaker embedding system. In the proposed system, the Barlow Twins objective function is calculated in the embedding layer and it is optimized jointly with the speaker classifier loss function. The experimental results on Fabiole corpus show 22 % relative gain in terms of EER in the clean environments and 18% improvement in the presence of noise with low SNR and reverberation.
Index Terms: Speaker recognition, ResNet, Barlow Twins, Robustness

1. Introduction

A speaker recognition system authenticates the user's identity from speech utterances. The state-of-the-art speaker recognition systems are mainly based on DNNs to extract a fixed-size compact representation from variable-length speech utterances known as speaker embedding or x-vector. The TDNN [1], CNN [2], ResNet [3], and VGGVox [3] speaker embedding systems are among widespread and successful architectures.

Although, the DNN-based speaker embedding systems have given a degree of robustness against acoustic noises, there is a significant degradation of their performance in the presence of background noise, reverberation and other variabilities [4] [5] [6]. Various approaches have been proposed to handle these variabilities in different parts of the system such as: signal level [7], feature level [8], speaker modeling level [9], x-vector level [6] and scoring technique level [10]. Addressing the variabilities at each step has its own advantage and disadvantages in terms of data, computational resources, efficiency, etc. In this paper we chose to make the ResNet-based speaker embedding system more robust against background noise and reverberation with a self-supervised objective function named Barlow Twins [11] . In the current work we worked on speaker modeling level. Because reducing the impact of noise and reverberation in higher levels is limited [12], having a noise robust speaker embedding system is highly demanded.

The goal of self-supervised learning (SSL) is to learn robust and invariant representation of the same data samples in the presence of different distortions (i.e. additivee noise and reverberation in our case). Several self-supervised learning methods are proposed for robust data representation [13] [14] . Among the proposed methods contrastive loss function is applied in robust speaker recognition system [3] and language adaptation [15]. Although the contrastive loss function has given promising results in domain adaption and robust speaker recognition, it has some limitations such as necessity for large batch size and the way of defining the negative pairs [16].

In this paper, we introduce the Barlow Twin objective function in the domain of speaker recognition systems. Barlow Twins is a self-supervised objective function that has two goals. Firstly, it increases the similarity between two versions of the same signal to give invariant representation. Secondly, it reduces the redundancy between different dimensions of x-vectors. The first goal makes the speaker representations more robust against variabilities and the second goal improves the discriminability of representations that improves the overall quality of the representations [11] [17]. The Barlow Twins objective function is integrated with the ResNet-based speaker embedding system. In the proposed system, the Barlow Twins objective function is jointly optimized with the Arc Margin SoftMax loss function. The Arc Margin SoftMax is obtained from the last layer of the ResNet system that classifies the speakers and the Barlow Twins function is calculated over the clean version and its noisy corresponding version of x-vectors extracted at the embedding layer of the ResNet network at each mini-batch.

In the following parts of this paper, firstly the related works are reviewed in section 2. The proposed method is described in section 3. The experiments setup is explained in section 4 and results are discussed in section 5.

2. Related works

Noise and reverberation are treated in the different parts of speaker recognition systems. In this section, we review works in the speaker modeling level (x-vector extractor), which are directly related to our work.

Some self-supervised learning methods are used for domain adaptation in speaker recognition systems. In a self-supervised approach, [**?**] the SoftMax loss function is trained with the contrastive loss. Because joint training of SoftMax loss function and contrastive function is intricate, they optimized the contrastive loss to fine-tune a network that is pretrained with SoftMax loss function. In [18] a domain adaptation technique is proposed that uses mean discrepancy distance (MMD) as a regularizer integrated with speaker embedding that performs the

Table 1: *The baseline ResNet-34 architecture.*

Layer name	Structure	Output
Input	–	$60 \times 400 \times 1$
Conv2D-1	3×3, Stride 1	$60 \times 400 \times 32$
ResNetBlock-1	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$, Stride 1	$60 \times 400 \times 32$
ResNetBlock-2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 4$, Stride 2	$30 \times 200 \times 64$
ResNetBlock-3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 6$, Stride 2	$15 \times 100 \times 128$
ResNetBlock-4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$, Stride 2	$8 \times 50 \times 256$
Pooling	–	8×256
Flatten	–	2048
Dense1	–	256
Dense2 (Softmax)	–	N
Total	–	–

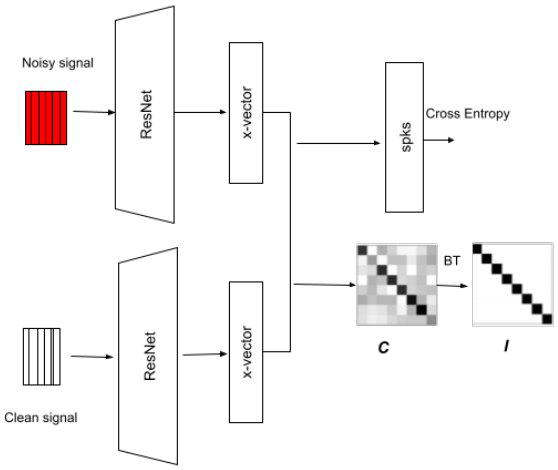

Figure 1: *Robust SR with Barlow Twins Loss function.*

adaptation between source and target domain. In this paper the proposed method is tested on language adaptation and its efficiency for noise and reverberation adaptation is not examined.

In another line of research, adversarial training is used to make the speaker embeddings more robust against domain mismatch. In [19] an adversarial strategy was proposed to make the speaker embedding more robust against noise. In the standard x-vector extractors, after the embedding layer, a DNN speaker classifier is optimized. In this work, a second classifier is trained adversarially to classify the type of noise. In another work, a GAN-based speaker embedding proposed that uses a binary discriminator to discriminate noisiness of the x-vector alongside the speaker recognition classifier [20]. The main deficiency of adversarial speaker embedding systems is the labels that should be used in the discriminator. Moreover, training the speaker embedding network in a manner that can not discriminate the type of noise or the noisiness of an x-vector doesn't guarantee that noisy x-vectors are close enough to their clean version.

To the best of our knowledge the current paper is first attempt of using Barlow Twins in speech processing applications in general and specifically in the domain of robust speaker recognition.

3. Proposed approach

3.1. Baseline system

The baseline embedding extractor used in this paper is a variant based on ResNet [21]. The ResNet model for extracting embeddings consists of three modules: a set of *ResNet Blocks*, a *statistics-level* layer, and *segment-level* representation layers.

- ResNet (Residual Network) uses stack of many Residual Blocks. A Residual Block is made up of two 2-dimensional convolutional Neural Networks (CNN) layers separated by a non-linearity (ReLU). The input of Residual Block is added to its output in order to constitute the input of the next Residual Block.
- The *statistics-level* component is an essential component to convert a variable length speech signal into a single fixed-dimensional vector. The statistics-level is composed of one layer: the statistics-pooling, which aggregates over frame-level output vectors of the DNN and computes their mean and standard deviation.
- The *segment-level* component maps the segment-level vector to speaker identities. The mean and standard deviation are concatenated together and forwarded to additional hidden layers and finally to softmax output layer.

The detailed topology of the used ResNet is shown in Table 1. Batch-norm and ReLU layers are not shown. The dimensions are (Frequency$\times$Channels$\times$Time). The input is comprised of 60 Mel scale filter banks from speech segments. During training we use a fixed segment length of 400 frames equals 4 seconds. The speaker ResNet system is trained with Arc Margin SoftMax loss function.

3.2. Barlow Twins system

In the Barlow Twins system, the baseline ResNet network accepts the clean version and its augmented version of the clean signal at each mini-batch. Therefore, in the embedding layer, we have pairs of clean and noisy embeddings that fed into the speaker classifier and the Barlow Twins objective function. The architecture of the proposed system is depicted in Fig. 1. The generator accepting noisy and clean signals are identical.

The Barlow Twins objective functions accepts two sets of inputs: z^X and z^Y are the mean centered normalized versions of clean and noisy x-vectors obtained from the embedding layer of ResNet system at each mini-batch.

The Barlow Twins function is defined as Eq.1:

$$L_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_{i!=j} (C_{ij})^2 \qquad (1)$$

where C is the correlation matrix between the output of noisy and clean x-vectors at each mini-batch and, i and j are x-vector dimensions. The C matrix is a square matrix; its dimension is equal to the size of the speaker embedding. The first term is called invariant term which tries to increase the correlation between two versions of noisy and clean x-vectors, in order to give an invariant x-vector for noisy and clean versions, the second term is called redundancy reduction which reduce the redundancy within the dimensions of the embedding, the λ coefficient indicates the importance of each term, and C_{ij} is defined as Eq.2:

$$C_{ij} = \frac{\sum_b z^X_{b,i} z^Y_{b,j}}{\sqrt{\sum_b (z^X_{b,i})^2} \sqrt{\sum_b (z^Y_{b,j})^2}} \quad (2)$$

the sum is performed over all the embeddings in given mini-batch, and b is the index of an embedding in the mini-batch.

In the proposed system, the Barlow Twins objective function is optimized jointly with the Arc Margin SoftMax loss function that is used in the speaker classifier. The final objective function is the summation of Arc Margin SoftMax loss and Barlows Twins objective function. Both functions are optimized with the same weight. In the end, the Barlow Twins objective function imposes on the correlation matrix to be an identity matrix (it maps the C matrix to I identity matrix shown in Fig. 1). The optimization of Barlow Twins brings the diagonal elements closer to 1 to have the invariant representations and imposes on off-diagonal elements of the correlation matrix to be close to 0.

Considering that a mini-bach is a matrix of D rows and B columns; where B is the size of the mini-batch and D is the size of an embedding, we can see that C_{ij} is the cosine between the vector lines of indices i ($z^X_{b,i}$) and j ($z^Y_{b,j}$) of the mini-batch. In this case, the Barlow Twins objective function consists of minimizing the distances between row vectors with the same indices and maximizing the distances between row vectors having different indices. In this sense, it is similar to the contrastive learning objective function. Indeed, it is the same formulation applied to the lines of the mini-batch in the case of the Barlow Twins and applied to the columns of the mini-batch in the case of contrastive learning. In the calculation of invariant term in Barlow Twins, each element comes from different speakers with different noises. It means that both noise and speaker variabilities are present. While in calculating the distance between the positive pairs in contrastive loss only the noise variability is taken into account because two samples come from the same speaker that are augmented with different noises.

4. Experiments setup

In this section the experiments setup including the used datasets, x-vector extractors and evaluation protocols are described.

4.1. Datasets

The datasets used in our paper are as follows:

- **Voxceleb2**. The Voxceleb2 [3] is used to train the x-vector extractors. There are 1.2m samples from 5,994 speakers. The clean version is augmented with Musan corpus and RIR files. The final version of training data included 5.9m samples.
- **Musan**. Musan is a music, speech, and noise corpus comprising 109 hours of speech data. All branches of Musan corpus are used for data augmentation to train x-vector extractors [22].
- **BBC Noise**. BBC Noise includes 16000 noise files, provided by BBC. These noises are used as artificial noise for evaluation protocols[1].
- **Fabiole1**. Fabiole1. is a French corpus that contains 7,000 files from 130 speakers [23]. We created an evaluation protocol from this dataset.
- **Robovox**. It is a French corpus collected from Robovox project (A mobile robot). Each recording in this corpus has 5 channels. The fifth channel is close microphone in which we considered it as clean and other channels are considered as noisy and reverberated[2]. The utterances are recorded in both open and closed spaces. The distance between speaker and microphones in far channels is between 1 and 3 meters.

[1] http://bbcsfx.acropolis.org.uk

4.2. x-vector extractors

Both baseline and Barlow Twins x-vector extractors are trained with Voxcleeb in 10,000 iterations. In another experiment, the Barlow Twins were used with a pretrained baseline system. In the last case, the Barlow Twins and Arc Margin SoftMax loss function are optimized together for 1,000 more iterations of the baseline system. The learning rate at the beginning of the training is set to 0.2 with weight decay equals 2.10^{-4}. The momentum is set to 0.9. The gradient descent optimizer is used. The size of the feature maps is 32, 64, 128, and 256 for the 4 ResNet blocks.

- **Baseline**. In the baseline system, the training samples are chosen randomly. The training data includes all clean files of Voxceleb and their augmented version with Musan Corpus, and reverberated with a pool of RIR files. [3]. Kaldi toolkit is used for data augmentation [24]. The SNR was chosen between 0 and 20. The batch size is set to 128. In this system only Arc Margin SoftMax loss function is optimized.
- **Barlow Twins**. In this system, a clean file from Voxceleb was chosen randomly. After that, its augmented version was chosen. Because we have two versions of each file at each mini-batch, we reduced the size of each mini-batch to 64. At the embedding layer, the Barlow Twins objective function is calculated and the proposed system was updated to minimize the summation of Barlow Twins and Arc Margin SoftMax functions. The λ varibale is set to 0.005 in Eq. 1. The λ is chosen experimentally.

4.3. Evaluation protocols

- **Fabiole**. In the first protocol, the Fabiole corpus is used. In this protocol 130 files (one file per speaker) are used as enrollment and 6,870 randomly chosen files are used for the test. In this protocol the BBC noise files are added to the clean signal with different SNRs from 0 to 15. In all cases the clean signal is used for enrollment. In this protocol the Kaldi toolkit is used to add noises to clean files.
- **Robovox**. In this protocol 26 files, one file per speaker, are used as the enrollment and 677 files are used as the test. The enrollment files are chosen from a close microphone with high quality but the test files are chosen from far microphones. The average length of speech utterances in this protocol is 22 seconds.

The details of both protocols are summarized in Table 2.

5. Results and discussions

In this section the obtained results are discussed. The results obtained from Fabiole protocol are depicted in Table 3. The BT column shows the results for a system in which Barlow Twins is

[2] https://robovox.univ-avignon.fr/

[3] http://www.openslr.org/resources/28/rirs_noises.zip

Table 2: *Experimental Protocols*

Protocol	Trials	Test	Enrolment
Fabiole	893k	6870	130
Robovox	17k	677	26

Table 3: *Fabiole Protocol (EER)*

SNR	Baseline	BT	Pre+BT
Clean	6.27	**4.87**	5.46
[0-5]	8.31	**6.81**	7.37
[5-10]	7.43	**5.86**	6.66
[10-15]	6.87	**5.48**	6.17

optimized from scratch with the speaker classifier. For example in a clean environment EER reduce from 6.27 to 4.87 which means 22% relative gain. In the case of low SNR between 0 and 5, we achieved an 18% relative gain of EER. The results for a case that Barlow Twins were used with a pretrained baseline system are presented in the last column. In this experiment, in all cases, we observed significant improvement of EER but the results for the training of Barlow Twins from scratch are better.

The results with the Robovox protocol are presented in Table 4. In the BT column, the results are shown for the case of joint optimization of both loss functions from scratch. In this experiment, we observed significant improvement for some channels but the behavior is not the same in all channels. The obtained results show that in the clean situation (i.e. channel 5) the Barlow Twins improves the performance. In other channels that are far and noisy, the results are paradoxical. Finally the results for an experiment that Barlow Twins adapts the pretrained baseline system are presented in the last column. In this case we observed improvement in all channels for example in clean environment there is 33% relative gain.

6. Conclusion

In this paper, the Barlow Twins objective function was introduced in the area of robust speaker recognition systems. The Barlow Twins objective function integrated with ResNet speaker embedding in order to achieve two goals: give an invariant representation for both clean and noisy versions of a speech signal, and reduce redundancy between different dimensions of the speaker embedding. We showed that the Barlows Twins objective function improves the performance of the speaker embedding system in both noisy and clean environments. The joint optimization of contrastive loss and Barlow Twins loss function in robust speaker recognition is a potential future work. In future work, the behavior of the Barlow Twins objective function in the presence of specific noises and with more data augmentation techniques will be studied.

7. Acknowledgements

This work was supported by the Robovox ANR-18-CE33-0014 project.

Table 4: *Robovox Protocols (EER)*

Channel	Baseline	BT	Pre+BT
1	4.28	4.28	**3.98**
2	4.72	4.43	**4.13**
3	4.38	4.57	**3.84**
4	5.76	5.90	**5.31**
5	2.21	1.92	**1.47**

8. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[2] D. Cai, Z. Cai, and M. Li, "Deep speaker embeddings with convolutional neural network on supervector for text-independent speaker recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1478–1482.

[3] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," vol. 60, 2020, p. 101027.

[4] M. Mohammadamini and D. Matrouf, "Data augmentation versus noise compensation for x-vector speaker recognition systems in noisy environments," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1–5.

[5] M. Mohammadamini, D. Matrouf, and P.-G. Noé, "Denoising x-vectors for Robust Speaker Recognition," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 75–80.

[6] M. Mohammadamini, D. Matrouf, J.-F. Bonastre, R. Serizel, S. Dowerah, and D. Jouvet, "Compensate multiple distortions for speaker recognition systems," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 141–145.

[7] J. G. Suwon Shon, Hao Tang, "Voiceid loss: Speech enhancement for speaker verification," in *INTERSPEECH*, 2019.

[8] O. Novotný, O. Plchot, O. Glembek, J. ernocký, and L. Burget, "Analysis of dnn speech signal enhancement for robust speaker recognition," *Computer Speech Language*, vol. 58, pp. 403–421, 2019.

[9] Y. Ma, K. A. Lee, V. Hautamäki, and H. Li, "Pl-eesr: Perceptual loss based end-to-end robust speaker representation extraction," in *ASRU*, 09 2021.

[10] Q. Wang, K. Okabe, K. A. Lee, and T. Koshinaka, "A generalized framework for domain adaptation of plda in speaker recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6619–6623.

[11] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 12 310–12 320. [Online]. Available: https://proceedings.mlr.press/v139/zbontar21a.html

[12] M. Mohammadamini, D. Matrouf, J.-F. Bonatsre, S. Dowerah, R. Serizel, and D. Jouvet, "Le comportement des systèmes de reconnaissance du locuteur de l'état de l'art face aux variabiliés acoustiques," Feb. 2022, working paper or preprint. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03564767

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119.

PMLR, 13–18 Jul 2020, pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[14] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," 2020. [Online]. Available: https://arxiv.org/abs/2006.07733

[15] V. Brignatz, J. Duret, D. Matrouf, and M. Rouvier, "Language adaptation for speaker recognition systems using contrastive learning," in *Speech and Computer*, A. Karpov and R. Potapova, Eds. Cham: Springer International Publishing, 2021, pp. 91–99.

[16] C. Wu, F. Wu, and Y. Huang, "Rethinking infonce: How many negative samples do you need?" 2021. [Online]. Available: https://arxiv.org/abs/2105.13003

[17] Y.-H. H. Tsai, S. Bai, L.-P. Morency, and R. Salakhutdinov, "A note on connecting barlow twins with negative-sample-free contrastive learning," 2021. [Online]. Available: https://arxiv.org/abs/2104.13712

[18] W. Lin, M.-M. Mak, N. Li, D. Su, and D. Yu, "Multi-level deep neural network adaptation for speaker verification using mmd and consistency regularization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6839–6843.

[19] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, "Training multi-task adversarial network for extracting noise-robust speaker embedding," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6196–6200.

[20] J. Zhou, T. Jiang, Q. Hong, and L. Li, "Extraction of noise-robust speaker embedding based on generative adversarial networks," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1641–1645.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[22] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[23] M. Ajili, J.-F. Bonastre, J. Kahn, S. Rossato, and G. Bernard, "FABIOLE, a speech database for forensic speaker comparison," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 726–733. [Online]. Available: https://aclanthology.org/L16-1115

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.