Autoencoder-based Semi-Supervised Curriculum Learning For Out-of-domain Speaker Verification

Siqi Zheng, Gang Liu, Hongbin Suo, Yun Lei

Machine Intelligence Technology, Alibaba Group

{zsq174630, g.liu, gaia.shb, yun.lei}@alibaba-inc.com

Abstract

This study aims to improve the performance of speaker verification system when no labeled out-of-domain data is available. An autoencoder-based semi-supervised curriculum learning scheme is proposed to automatically cluster unlabeled data and iteratively update the corpus during training. This new training scheme allows us to (1) progressively expand the size of training corpus by utilizing unlabeled data and correcting previous labels at run-time; and (2) improve robustness when generalizing to multiple conditions, such as out-of-domain and text-independent speaker verification tasks. It is also discovered that a denoising autoencoder can significantly enhance the clustering accuracy when it is trained on carefully-selected subset of speakers. Our experimental results show a relative reduction of $30\% - 50\%$ in EER compared to the baseline.

Index Terms: Speaker Verification, Semi-Supervised Learning, Curriculum Learning, Denoising Autoencoder, SSCL

1. Introduction

As smart speakers such as Tmall Genie, Amazon Echo, and Google Home become popular, a robust speaker verification system tailored to the needs of these family devices has become increasingly important. With speaker verification system, smart speakers are able to recognize the identity of a user via voice and further to provide customized service and identity verification for security purpose [1]. However, acquiring large-scale labeled data for training on smart speakers is challenging. Unlike mobile phone devices, most of which are used by single user, these smart speakers are designed to be shared among multiple users, which means there is no simple one-to-one mapping between device-generated data and user. Traditionally, the common practice in industry is to record some predefined prompts in a studio or simulation environment, or manually label the speaker identities of the online data in order to harvest an ideal training corpus for machine learning algorithm.

As smart speakers become more and more commonly acceptable as family virtual assistants, a number of different generations and models have emerged to meet different consumers' needs. Often, the acoustic data collected from various models of smart speakers can be quite different. For example, model C1 of Tmall Genie has a 2-microphone array while model X1 of Tmall Genie operates on a 6-microphone array. This poses new challenge to speaker verification. Since collecting a large set of labeled speakers for training is labor-intensive and time-consuming, an efficient data leverage method is critical to keep up with the fast pace of releasing new models of smart speakers.

Another constraint in the industry is that some legal regulations require users' direct authorized consents for their recordings to be accessed and used for manual labeling. Collecting consents from a large number of speakers is impractical. Therefore, a novel scheme that can learn from unlabeled data and update its training corpus and model parameters at run-time is appealing.

Curriculum learning is first formalized by Bengio et al. and refers to the concept where machines start by learning simple structures and gradually improve to learn more complex ones [2]. Bengio et al. show that this "start-small-and-gradually-increase-the-difficulty" strategy helps lead to faster convergence and find better local minima of non-convex problems.

Motivated by this idea, Ranjan et al. proposed a curriculum learning based i-Vector PLDA approach to improve robustness on noisy data [3]. Training data are split into subsets and gradually included into the PLDA training procedure based on their difficulties. An LSTM-based speaker verification system is introduced in [4] where the parameters are optimized via a curriculum learning procedure. The network starts by learning from text-dependent samples before gradually introducing text-independent and noisy samples. However, both of the works are only suitable when labeled data are readily accessible for training (e.g. mobile phone users). In cases when collecting a large-scale labeled data is impractical, we have to either rely on limited labeled data, or take advantage of the large quantity of unlabeled data.

Denoising Autoencoders (DAE) have been widely explored in the field of speech signal processing [5][6][7][8][9]. [7] and [8] reported the effectiveness of DAE on dereverberation and distant-talking speech recognition. [10] investigated the performance of DAE on unsupervised domain adaption for speech emotion recognition. Pekhovsky et al. presented a detailed empirical investigation of the speaker verification system based on denoising autoencoder in the i-vector space[6]. Since reverberation, distant-talking, and domain variation are among the most common interfering factors for speaker verification on smart home speakers, we are intrigued to investigate how denoising autoencoders would improve the accuracy of unsupervised clustering on the diversified smart speaker data.

In this study an autoencoder-based semi-supervised curriculum learning approach is proposed to quickly adapt speaker verification system to the unseen new domain in which no labeled data is available. This training scheme enables us to tap the tremendous unlabeled data, which quickly increases the size of training corpus and results in significant performance improvement when compared to the baseline trained purely with labeled data. Our deep neural nets are bootstrapped by learning from a small amount of clean, labeled in-domain data from model X1, all speaking similar triggering keyword phrases ("Hello Tmall!" or "Tmall Genie!" in mandarin). As the training progresses, we gradually increase the complexity of training data by adding unlabeled utterances from other smart speaker model types, as well as text-independent utterances(e.g. "Tmall Genie, play some music"). The utterance labels are automatically generated by running unsupervised agglomerative clustering on speaker embeddings extracted from the intermediate

results at different epochs during the training process. At each epoch, denoising autoencoders are also re-trained and updated simultaneously. The data set used for training denoising autoencoders is a subset of the entire corpus, which involves only speakers that are explicitly enrolled. It will be demonstrated later in this study that the proposed unsupervised clustering approach allows for quick expansion of the size of training set, which in turn makes it possible to leverage the discriminating power of deep neural nets.

It has been shown in theory by [11] that the learned model can benefit from unlabeled data. It is also mentioned that additional information of the unlabeled data decreases as classes overlap. Therefore, it is essential that speaker classes are well-separated. In order to achieve this, we collect data that are grouped based on the device serial number (SN) of each Tmall Genie smart speaker. It is assumed that clustering on these home-based virtual assistants will end up in only a small number of clusters, and acoustic characteristics among clusters are significantly different due to the demographic structure. Our preliminary investigation on a random sample of 100 online devices further validates our assumption. As shown in Table 1, 91% of Tmall Genie speakers have no more than 2 users - most likely a male and a female. And a predominant number of Tmall Genies have no more than 3 users - a male, a female, and a child. The significant gender and age difference among users of the same device underpin our belief that this semi-supervised curriculum learning based training scheme will provide us with a large data set of promising quality.

# of users	Percentage
1	70%
$\leq$2	91%
$\leq$3	98%

Table 1: *Number of users from 100 online devices and their distribution.*

2. System Description

2.1. Baseline

In the baseline system the speaker embedding extracted from a time-delay neural network (TDNN) is used as speaker representation vector[12]. The TDNN is trained with only labeled data from Tmall Genie X1. The average length of the utterances is 1.7 seconds. All of the utterances are text-dependent - repeating the same wake-up keywords of Tmall Genie such as "Tmall Genie!" and "Hello, Tmall!". With the configuration of 20ms frame length and 10ms frame shift, a 60-dimension feature vector is extracted for each frame, including 20-dimension MFCC, deltas and delta-deltas.

The TDNN is trained with a cross entropy loss function and ReLU activation function. At the input layer we splice together five consecutive frames around each frame. For the next two layers we splice together the inputs from the previous layer at time-step 2 and 3, respectively. Hence for each frame, we obtain a context of 15 frames around it. After aggregating frame-level outputs using a statistics pooling layer, the network is followed by several segment-level layers before ending with a softmax layer. Speaker embeddings are extracted from one of the affine segment layers.

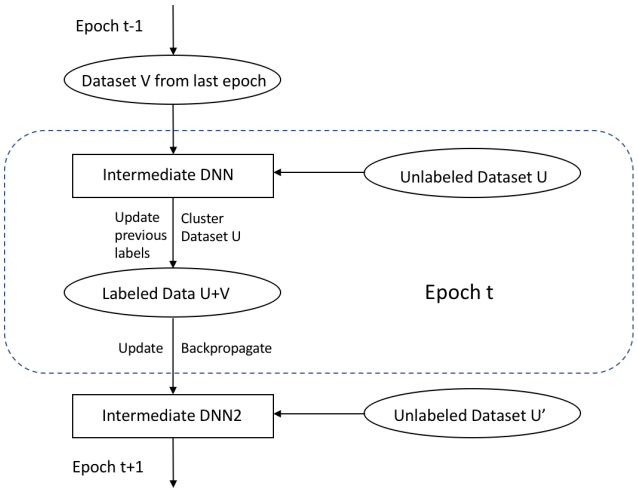

Figure 1: *At epoch t, new unlabeled dataset U is scored and clustered by the intermediate DNN from previous epoch. Labels from previous dataset V are corrected before combining new dataset U. The newly labeled dataset U+V are then utilized to train and update DNN.*

2.2. Semi-Supervised Curriculum Learning

In this section we describe the training scheme of Semi-Supervised Curriculum Learning (SSCL). We demonstrate how it helps us to improve out-of-domain performance, and generalize to text-independent scenarios when the amount of labeled data is limited and the conditions of training data are constrained, i.e. in-domain and text-dependent.

SSCL bootstraps the training with a small number of text-dependent, recorded data from Tmall Genie X1 speaker. When the validation accuracy reaches a reasonable empirical threshold, SSCL repeatedly performs unsupervised speaker clustering using intermediate deep neural nets at the end of previous epoch. At every epoch, the speaker embedding extracted from the intermediate deep neural nets will propagate through an intermediate denoising autoencoder. The outputs of denoising autoencoder are used for clustering. We will elaborate the training of denoising autoencoder in the last part of this section.

The bottom-up agglomerative clustering approach as described in [13][14] is used to cluster speakers at each epoch. Each of the utterances are initialized as its own cluster. At each step, the mean similarity scores between every cluster and all other speaker embedding clusters are calculated. Two clusters with the highest mean similarity score are then merged. We repeat this process until no mean similarity scores between any two clusters meet the stopping threshold.

The results of agglomerative clustering are added into the training data pool for the next epoch. By doing this we can quickly expand the size of the training data. Every time SSCL finishes an epoch with larger training data set, it acquires a better discriminating ability than its predecessor. Hence SSCL also iteratively corrects the previous labels with better clustering accuracy. Data with corrected labels are fed back into the training corpus, which will reinforce SSCL's learning ability later. The number of nodes in the output layer of the neural network is also restructured to match the updated number of labels in the new training corpus.

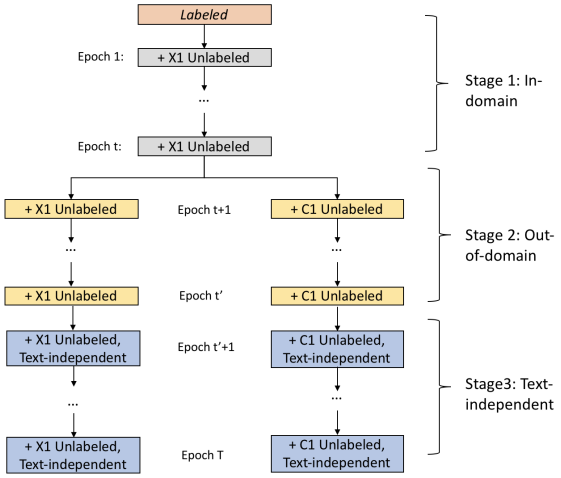

Figure 2: *Three major stages of the SSCL. Start with labeled data. Train with text-dependent X1 unlabeled data in Stage 1, text-dependent unlabeled data from X1 and C1 in Stage 2. Finally, train with independent data in Stage 3.*

Besides harvesting new utterances and correcting previous labels, SSCL also removes outliers and noisy data every time when the stopping criteria of unsupervised clustering is met. Utterances that do not belong to any of the clusters or clusters with too few utterances are removed from the training corpus for the next epoch of training.

For the next step on the curriculum, we increase the complexity of the training data by introducing out-of-domain data. Utterances from other smart speaker model (Tmall Genie C1 model in this study) are fed to the learning scheme. Without having any labeled data for the new domain, we follow a similar approach - iteratively update the neural nets and cluster speakers using TDNN trained from previous step. As the number of speakers from new domain increases, speakers and utterances from old domain (data from Tmall Genie X1) are gradually removed from the training corpus. This way SSCL progressively adapts itself to the new domain.

After SSCL has learned to handle the complexity from a different domain, we try to enable it to master more difficult contents on the curriculum. We want SSCL to generalize to text-independent case. As mentioned earlier, the labeled data set consists of only triggering utterances such as "Tmall Genie". In the next learning stage, SSCL is fed with text-independent data and triggering keywords followed by random speech (for example, "Tmall Genie, What's the weather tomorrow.").

2.3. Implicit vs. Explicit Enrollment

Before discussing the effect of including a denoising autoencoder in our system, we first describe the concept of implicit and explicit enrollment, two types of enrollment approaches that are becoming more and more common in the scenario of smart speakers.

Explicitly enrolled speakers refer to users who take the initiatives to enroll their voices by following the smart speaker's setup guidance process.

Implicitly enrolled speakers are those who never intentionally set up their voice accounts but as the smart speakers accumulate their recordings, the speaker identification algorithm implicitly enroll their voices using the same unsupervised clustering approach mentioned in the SSCL.

Explicitly enrolled utterances are deemed to be more reliable and often be of higher quality. When users consciously follow the setup guidance to enroll their voices, they are facing the smart speakers in a close distance. Therefore, an utterance marked as "explicitly enrolled" is a indication that it is a near-field utterance with relatively reliable quality.

On the other hand, implicitly enrolled utterances are collected from users' daily interactions with smart speakers. They are mostly far-field utterances with the presence of background noises. Therefore, pairing up an explicit enrollment with implicit enrollment from the same user creates a near-field-to-far-field mapping to be used for training denoising autoencoder.

Our baseline system is trained on labeled utterances from explicitly enrolled speakers data only. For the SSCL training both implicitly and explicitly enrolled speaker data are used in unsupervised clustering.

2.4. Denoising Autoencoder

Our denoising autoencoder is trained to project embeddings from far-field space onto near-field space. We rely on the assumption that enrolled utterances collected from a user's initiative action are near-field and can be safely deemed to be from the same speaker. To ensure the quality of training inputs, a simple score comparison among enrolled utterances of the same speaker is performed to validate our assumption.

The mean of all enrolled embeddings of a user is formalized as the target feature of our DAE. All other non-enrolled speaker embeddings of that specific user are treated as the input features for DAE training.

The dimension of input and output layer of our denoising autoencoder is 512. It contains three hidden layers, each of which has a dimension of 1024. At each epoch during SSCL, all speaker embeddings extracted from TDNN will go through a forward-propagation via DAE. The output of DAE is used for agglomerative clustering.

The results of clustering in turn benefit DAE by updating the explicit enrollment corpus for the next epoch of training - new speakers are added, previous labels are corrected, and noisy cases are removed. In this manner, the parameters of denoising autoencoder are trained and updated alongside the parameters of main deep neural network.

As SSCL enters the out-of-domain stage and text-independent stage on the curriculum (stage 3 in Figure 2), it updates the training corpus for DAE in a similar way. Subsets of explicitly enrolled speakers are selected from out-of-domain and text-independent utterances, respectively.

3. Experiment

3.1. Corpus

The labeled data set consists of 220,000 utterances from 5000 X1 speakers, recorded in a simulation environment. All of the utterances are clean, text-dependent with the same triggering phrase.

The total pool of unlabeled data consists of 22 million utterances, collecting from over 100,000 online devices, from both X1 and C1. The size of the corpus keeps increasing at run-time and by the last epoch of SSCL, all 22 million utterances have been used for training - except those have been determined as "bad data" and discarded. Texts range from fixed triggering phrases to completely random mumbles.

The test set contains 300 speakers from X1 and 161 speakers from C1. Each of the speakers are asked to record 20-40 utterances on specified simulation devices.

3.2. Trials & Measurements

Trials are designed to reproduce the real world scenarios as close as possible. Both target and non-target trials are contrived based on device IDs. There will be no cross-device trials as they are unfeasible during actual usage. For example, the voice registered on a home smart speaker can only be accessed through the same device. This amplifies the difficulty of the task. In most traditional speaker verification tasks, utterances from different speakers are recorded on different devices. The speaker features extracted are, in fact, mixtures of both speakers and devices information. The effect of device difference intensifies the gap between utterances from different speakers, which in turn makes them easier to be discriminated.

In addition, confusion errors are used to measure performance. Confusion error is formalized in [15][16] and refers to the condition where the score exceeds the threshold but the class associated with the maximum score is not the correct enrolled speaker. In this study, confusion error rate is defined as

$$Conf.err. = \frac{\sum_i f(i)g(i)}{\sum_i f(i)}$$

where

$$f(i) = \begin{cases} 1 & s_i > Thr \\ 0 & s_i \leq Thr \end{cases}$$

and

$$g(i) = \begin{cases} 1 & s_i < max(s_{ij}) \\ 0 & s_i = max(s_{ij}) \end{cases}$$

where s_i is the score of test utterance i, Thr is the verification threshold, and $max(s_{ij})$ refers to the maximum score of test utterance i against all enrolled utterances on device j to which test utterance i belongs. To sum up, confusion error rate measures, among all test trials whose scores exceed the threshold, the proportion of test utterances whose maximum score mistakenly associates to another enrolled speaker.

4. Results

As shown in table 2, the baseline reports an EER of 10.86%. SSCL-X_t stands for the intermediate results at the tth epoch in Figure 1. SSCL-X_t is trained on a corpus whose size has been expanded to over ten thousands speakers and two millions of utterances. Similarly, SSCL-X_T is the result from epoch T.

	EER	*Conf. err.*
Baseline	10.86%	3.81%
SSCL-X_t	7.80%	3.08%
SSCL-$X_{t'}$	7.13%	2.51%
SSCL-DAE-X_t	6.93%	2.51%
SSCL-DAE-$X_{t'}$	**6.28%**	**2.20%**

Table 2: *Performance Comparison on Model Type X1*

In order to adapt our neural net to the out-of-domain (C1), the SSCL-X_t is used in the pretrained model to initialize SSCL training on out-of-domain data. As shown in Table 3, SSCL-X_t has achieved a relative reduction of 24% in EER, without having any training data from C1. After including DAE we can achieve a 5.44% EER and 1.87% confusion error rate.

	EER	*Conf. err.*
Baseline	9.08%	3.45%
SSCL-X_t	6.94%	2.78%
SSCL-C_n	**5.97%**	**2.02%**
SSCL-DAE-X_t	6.31%	2.68%
SSCL-DAE-C_n	**5.44%**	**1.87%**

Table 3: *Performance Comparison on Model Type C1*

Table 4 summarizes the performance of our neural network when the total number of utterances in the training pool reached a certain level. The field "# of speakers" indicates the inferred number of speakers resulted from unsupervised clustering.

# of utts(millions)	Type	# of speakers	EER
0.22M	Supervised	5,000	9.08%
2M	Unsupervised	17,000	7.82%
10M	Unsupervised	100,000	6.94%
22M	Unsupervised	178,000	**5.97%**

Table 4: *Performance Comparison on different sizes of training corpus*

The improvement on text-independent task is significant. The intermediate neural net SSCL-$X_{t'}$ is used to initialize the training of text-independent stage. As Table 4 indicates, the final output, SSCL-DAE-X_T, achieves more than 50% reduction in EER compared to the baseline.

	EER	*Conf. err.*
Baseline	7.86%	2.42%
SSCL-X_T	3.92%	1.57%
SSCL-DAE-X_T	**3.59%**	**1.43%**

Table 5: *Performance Comparison on Text-independent Utterances from Tmall Genie Model Type X1*

5. Conclusion

This study provides a semi-supervised training scheme to address one of the major industrial challenges of speaker verification - the scarcity of out-of-domain labeled data, despite the surplus of unlabeled data. Our experiments suggest that the potential harm incurred by wrong label prediction can be ameliorated by tremendously expanding the size of training data and iteratively correcting the labels with better intermediate models and denoising autoencoders. It allows our speaker verification system to quickly adapt to the release of any new models of smart speakers. Finally, as next step, we plan to extend the exploration beyond smart speakers to other fields in the industry where labeled speakers are scarce but unlabeled data are abundant.

6. Acknowledgement

We would like to thank Yao Sun, Wanlin Wang, and Ping Zhang for their contributions in providing the corpus that makes this research possible.

7. References

[1] G. Liu, Q. Qian, Z. Wang, Q. Zhao, T. Wang, H. Li, J. Xue, S. Zhu, R. Jin, and T. Zhao, "The opensesame nist 2016 speaker recognition evaluation system," in *Proc. Interspeech 2017*, 2017, pp. 2854–2858. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-997

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18*, 2009, pp. 41–48. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553380

[3] S. Ranjan and J. H. L. Hansen, "Curriculum learning based approaches for noise robust speaker recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 26, no. 1, pp. 197–210, 2018. [Online]. Available: https://doi.org/10.1109/TASLP.2017.2765832

[4] E. Marchi, S. Shum, K. Hwang, S. Kajarekar, S. Sigtia, H. Richards, R. Haynes, Y. Kim, and J. Bridle, "Generalised discriminative transform via curriculum learning for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20*, 2018, pp. 5324–5328. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461296

[5] S. Novoselov, T. Pekhovsky, O. Kudashev, V. S. Mendelev, and A. Prudnikov, "Non-linear PLDA for i-vector speaker verification," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10,*, 2015, pp. 214–218. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2015/i15_0214.html

[6] T. Pekhovsky, S. Novoselov, A. Sholohov, and O. Kudashev, "On autoencoders in the i-vector space for speaker recognition," in *Odyssey 2016: The Speaker and Language Recognition Workshop, Bilbao, Spain, June 21-24*, 2016, pp. 217–224. [Online]. Available: https://doi.org/10.21437/Odyssey.2016-31

[7] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29*, 2013, pp. 3512–3516. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2013/i13_3512.html

[8] Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, and M. Iwahashi, "Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification," *EURASIP J. Audio, Speech and Music Processing*, vol. 2015, p. 12, 2015. [Online]. Available: https://doi.org/10.1186/s13636-015-0056-7

[9] H. Xing, G. Liu, and J. H. Hansen, "Frequency offset correction in single sideband (ssb) speech by deep neural network for speaker verification," in *Proc. Interspeech 2015*, 2015.

[10] J. Deng, Z. Zhang, F. Eyben, and B. W. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, 2014. [Online]. Available: https://doi.org/10.1109/LSP.2014.2324759

[11] T. J. Oneill, "Normal discrimination with unclassified observations," *Journal of the American Statistical Association*, vol. 73, no. 364, p. 821, 1978.

[12] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24*, 2017, pp. 999–1003. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0620.html

[13] D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary clustering," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, 2006, pp. 554–560. [Online]. Available: https://doi.org/10.1145/1150402.1150467

[14] F. Murtagh and P. Contreras, "Methods of hierarchical clustering," *CoRR*, vol. abs/1105.0121, 2011. [Online]. Available: http://arxiv.org/abs/1105.0121

[15] E. Singer and D. A. Reynolds, "Analysis of multitarget detection for speaker and language recognition," in *ODYSSEY 2004 - The Speaker and Language Recognition Workshop, Toledo, Spain, May 31 - June 3*, 2004, pp. 301–308. [Online]. Available: http://www.isca-speech.org/archive_open/odyssey_04/ody4_301.html

[16] S. Zheng, J. Wang, J. Xiao, W. Hsu, and J. Glass, "A noise-robust self-adaptive multitarget speaker detection system," in *24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24*, 2018, pp. 1068–1072. [Online]. Available: https://doi.org/10.1109/ICPR.2018.8545395