INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Anonymous speaker clusters: Making distinctions between anonymised speech recordings with clustering interface

Benjamin O'Brien[1], *Natalia Tomashenko*[2], *Anaïs Chanclu*[2], *Jean-François Bonastre*[2]

[1] Aix-Marseille Univ, CNRS, LPL, UMR 7309, France
[2] Laboratoire Informatique d'Avignon, Université d'Avignon, Avignon, France
benjamin.o-brien@univ-amu.fr

Abstract

Our study examined the performance of evaluators tasked to group natural and anonymised speech recordings into clusters based on their perceived similarities. Speech stimuli were selected from the VCTK corpus; two systems developed for the VoicePrivacy 2020 Challenge were used for anonymisation. The Baseline-1 ($B1$) system was developed by using x-vectors and neural waveform models, while the Baseline-2 ($B2$) system relied on digital-signal-processing techniques. 74 evaluators completed three trials composed of 16 recordings with either natural or anonymised speech generated from a single system. F-measure and cluster purity metrics were used to assess evaluator accuracy. Probabilistic linear discriminant analysis (PLDA) scores from an automatic speaker verification system were generated to quantify similarity between recordings and used to correlate subjective results. Our findings showed that non-native English speaking evaluators significantly lowered their F-measure means when presented anonymised recordings. We observed no significance for cluster purity. Pearson correlation procedures revealed that PLDA scores generated from natural and $B2$-anonymised speech recordings correlated positively to F-measure and cluster purity metrics. These findings show evaluators were able to use the interface to cluster natural and anonymised speech recordings and suggest anonymisation systems modelled like B1 are more effective at suppressing identifiable speech characteristics.

Index Terms: privacy, anonymisation, speech synthesis, speaker identification, clustering, subjective evaluation

1. Introduction

Speech data contains various types of personal information about speakers, which can be revealed by human listeners or by automated systems [1]. This information includes, among other things, speaker identity, age, gender, ethnic origin, health or emotional state, political orientations, and religious beliefs [2]. In recent years, there has been a growing interest in privacy preservation solutions for speech technology. To promote the development of privacy preservation techniques for speech, the *VoicePrivacy initiative* has recently been introduced in [3], which focused on the voice anonymisation task.

The goal of voice anonymisation is to suppress personally identifiable information within the speech signal, while maintaining all other characteristics [3]. In particular, preserving the linguistic content of anonymised speech data and speech naturalness is essential. Before data processing and publication, speakers apply an anonymisation method to their utterances in order to hide their identity. Ideally anonymised utterances should sound as if they were produced by another speaker, which may be an artificial voice not belonging to any existing speaker.

To evaluate the effectiveness of an anonymisation algorithm, various subjective and objective methods have been proposed in recent works [3, 4, 5, 6, 7]. Evaluation methodologies depend on the privacy preservation scenarios. In this work, we focused on subjective evaluation and considered the scenario of an attacker attempting to access a set of original and anonymised utterances from multiple speakers. Our goal was to examine performance accuracy when evaluators were tasked to link natural and anonymised speech recordings.

Several methods are often used to examine the speaker identification performance accuracy. In the domain of forensic linguistics, speech recording lineups are often used to gather evidence, however, there have been criticisms regarding the effectiveness of the practice [8], as people rely on and use sensory information differently. A much simpler task involves a binary approach, where evaluators are tasked to determine whether two speech recordings belong to the same speaker or not. However, this method introduces memory bias, which can raise questions to any reported findings. By adding the criteria of comparing natural and anonymised speech, an alternative method might prove to be better suited to evaluate the effectiveness of anonymisation systems.

A major goal of our study was to develop and investigate a subjective evaluation methodology for assessing anonymisation algorithms. To do so, we required an interface that provided a platform for evaluators to (re-)listen to natural and anonymised speech recordings, so that they might link them to similar speakers. Having reported successful findings with a *clustering* method [9], we theorised that this approach would allow evaluators to personalise their engagements with speech materials and organise them by their perceived similarities. By comparing the performance of evaluators who evaluated natural and anonymised speech recordings between different systems, we might gain insight as to whether they were able to link similar speakers. Alternatively we might assess whether one system was more or less effective at suppressing identifiable speech characteristics, as it influenced speaker identification performance. It was also of interest to assess evaluator responses to natural and anonymised speech in relation to objective scores representing the likelihood that the clustered speech recordings belonged to the same speaker. By measuring the relationship between evaluator responses and the similarity of speech recordings, our goal was to examine more closely acoustic-perceptual correlates.

2. Method

2.1. Stimuli

Speech recordings were taken from the *VCTK-test (common)* dataset of the VoicePrivacy challenge [2, 10], which is com-

http://dx.doi.org/10.21437/Interspeech.2021-1588

posed of 700 speech recordings read by native-English speakers (female: 346; male: 354). Speech segments (*utterances*) #1-24 of 15 female speakers and 15 male speakers were selected. For each gender, the speakers were randomly divided into target speakers (9) and distractors (6). To standardise stimuli duration, 700 speech recordings were reduced to a maximum of 3 s (mean duration 2.915 $\pm$ 0.3 s). All recordings were then normalized to 0 dB. Section 2.3.3 describes our development of anonymisation systems used in the study.

2.2. Evaluators

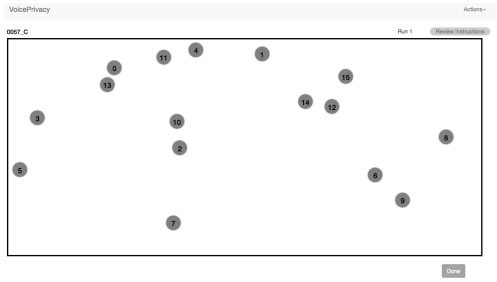
In total, 74 evaluators (26 female and 48 male) participated in the study. All evaluators reported good hearing. 29 were native-English speakers. The non-native English-speaking evaluators were either bilingual or held a high-level of English. The majority was Francophone (39).

2.3. Procedure

2.3.1. Task

Evaluators were tasked to listen to speech recordings and group them into clusters according to the perceived speaker voice similarities. Evaluators used a custom interface developed at Laboratoire Informatique d'Avignon (Avignon Université), which was accessible from a standard web browser. Figure 1 illustrates the cluster interface used by evaluators. Evaluators were provided with the instructions[1] and were encouraged to use personal headphones.

Figure 1: *An image of the cluster interface used by evaluators*



2.3.2. Trial design

Each evaluator completed three trials: 1 control and 2 evaluation trials (processed in a random order). Each trial included 16 different speech recordings, which were divided into 3 target speakers and 1 distractor speaker. Over the three trials, evaluators only encountered unique speakers, and all speakers had the same gender. Target speakers were allocated 2 to 6 utterances, while the distractor speaker was given 1 utterance. The control trial was composed of natural speech recordings, and its purpose was to assess a baseline clustering performance. The goal of evaluation trials was to assess evaluator performance when linking anonymised speech recordings with natural speech recordings. As a result, half of the 16 utterances were anonymised, including the distractor speaker, which was always anonymised. All anonymised speech recordings in a trial were processed by the same anonymisation system.

[1] https://demo-lia.univ-avignon.fr/voiceprivacy/instructions/

2.3.3. Anonymisation systems

In this paper, we considered two different anonymisation systems used as baselines for the VoicePrivacy 2020 Challenge[2][3].

The primary baseline [3], denoted as $B1$, was inspired from [7]. It is based on anonymisation using x-vectors and neural waveform models to synthesize anonymised speech. Its development required three steps. First, speaker x-vectors, pitch, and linguistic features were extracted from the speech signal. Then, x-vector anonymisation was performed, where the original speaker x-vector was replaced by a new anonymised x-vector. Finally, the original linguistic features, pitch, and anonymised x-vectors were used to synthesise anonymised speech by means of neural acoustic and waveform models. More details on B1 development are available in [2, 3, 11].

The secondary baseline [2], denoted as $B2$, is based upon signal processing techniques. In contrast to $B1$, it does not does not require any training data and is based on vocal tract filter transformations. $B2$ anonymisation algorithm is applied on the frame level and exploits the McAdams coefficient [12] to perform anonymisation by shifting the pole positions derived from the linear predictive coding (LPC) analysis of speech signals. For more details, please refer to [2, 13].

2.4. Data processing

2.4.1. Metrics

There are several methods used to evaluate clustering performance [14][15]. Figure 2 diagrams the general evaluation scheme developed in our study: (*i*) stimuli are selected for an evaluation trial; (*ii*) stimuli are randomly distributed on the interface; (*iii*) a evaluator arranges the stimuli into different clusters; and (*iv*) for each cluster, a principal speaker is identified. For (*iv*), however, we identified two principal ways to identify the cluster centroid (*proto-speaker*), which, as a result, can affect evaluator performance assessment.

For the first method we identified the mode speaker in a cluster as the proto-speaker, which permits the possibility of multiple clusters being linked to the same speaker. In the case where multiple modes were identified, one was selected randomly. For each cluster we calculated the macro-average F-measure (1), which evaluates both precision and recall[3]. For each cluster we identified the number of true and false positives followed by the number of false negatives in the trial and calculated F_1 as:

$$F_1 = \frac{tp}{tp + \frac{1}{2}(fp + fn)} \tag{1}$$

where tp is *true positive*, fp is *false positive*, fn is *false negative*.

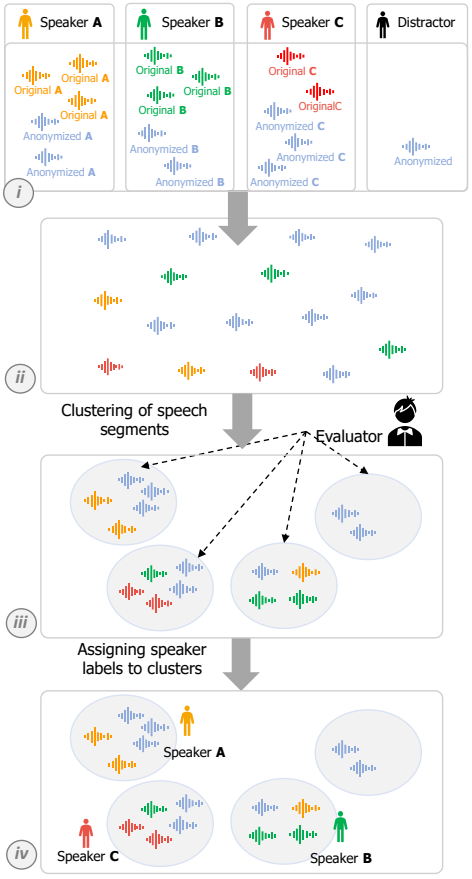
As an alternative method, we proposed the *cluster purity* metric, which identifies a different speaker to each cluster in a trial (2). Unlike F_1, purity focuses only on maximising the total number of true positive responses per cluster. Purity values range between 0 and 1 (perfect clustering). We define purity as:

$$purity(M) = \max_k \frac{1}{N} \sum_{m \in M} m \cap d_m^k \tag{2}$$

[2] Baselines are available at https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

[3] Precision is the proportion of total true positive responses to total positive responses. Recall is the proportion of total true positive responses to the total correct responses.

Figure 2: *Scheme for evaluation trial*



where M is a trial, m is a cluster in M, d^k is the different combinations of unique speakers assigned to each cluster in M, and N is the number of speech recordings in the trial.

It was important to assess whether other performance factors were affected by the presence of anonymised speech recordings. Thus, in addition, we evaluated the mean number of times evaluators listened to a speech recording (*listening count*).

2.4.2. *Comparison with objective evaluation*

In this section, we compare the subjective evaluation results and the objective results obtained by means of an automatic speaker verification (ASV) system. For this purpose, we used an ASV system, which relies on x-vector speaker embeddings and probabilistic linear discriminant analysis (PLDA) [16]. The ASV system was trained on the *LibriSpeech train-clean-360* dataset [17] as described in [2]. The ASV model was used to obtain PLDA scores [18] for evaluated pairs of speech segments, where s_a, s_b denote a pair of utterances. We computed the PLDA scores as log-likelihood ratio values between corresponding x-vectors x_a, x_b as:

$$\text{PLDA}(s_a, s_b) = \log \frac{P(x_a, x_b|\mathcal{H}_{\text{same}})}{P(x_a, x_b|\mathcal{H}_{\text{different}})} \quad (3)$$

where $\mathcal{H}_{\text{same}}$ and $\mathcal{H}_{\text{different}}$ are respectively the *same speaker* and *different speakers* hypotheses.

To calculate the objective score for each cluster, we applied the same methods used to identify the cluster proto-speaker for calculating F_1 and purity metrics. Once identified we selected the maximum mean difference between it and other speech recordings in a cluster.

2.5. Preliminary results

Normal distribution functions were fitted to evaluator mean duration to complete the trials. Two evaluators were excluded from further analysis, as their means were greater than three standard deviations from their group means. $B1$ (33) and $B2$ (39) evaluators completed the three trials in 343.47 $\pm$ 189.8 s and 323.02 $\pm$ 132.48 s, respectively.

The difference between performance during the control trial and the average performance during evaluation trials was used to measure the effects of anonymisation systems. For all outcome variables, mixed ANOVA (α = 0.05) procedures were carried out with the trial anonymisation system ($B1$, $B2$) and language (native or non-native English speaking evaluators) as between-subjects factors and speech recording gender (female, male) as within-subject factors. Where main effects were detected, post-hoc Bonferroni-adjusted t-tests were carried out.

3. Results

3.1. Difference between control and evaluation trials

We found a main effect for mean F_1 difference on language $F_{1,64}$ = 6.5, p $<$ 0.05, η_p^2 = 0.09, but no effects on system nor speech recording gender, p $>$ 0.05. $B1$ evaluators had a greater mean F_1 differences (0.24 $\pm$ 0.02) in comparison to $B2$ evaluators (0.21 $\pm$ 0.02). Post-hoc t-tests showed non-native English speaking evaluators were more affected by linking natural and anonymised speech recordings (0.26 $\pm$ 0.02) in comparison to native English speaking evaluators (0.19 $\pm$ 0.022). We observed no differences between means for female (0.21 $\pm$ 0.02) and male speech recordings (0.24 $\pm$ 0.02).

We found no significant main effects or interactions on mean purity difference, p $>$ 0.05. $B1$ evaluators were slightly more affected (0.13 $\pm$ 0.03) than $B2$ evaluators (0.14 $\pm$ 0.02). Non-native English speaking evaluators were similarly affected (0.13 $\pm$ 0.02) than native English speaking evaluators (0.13 $\pm$ 0.03). We observed a slight difference when evaluators were presented female speech recordings (0.13 $\pm$ 0.02) as compared to male speech recordings (0.14 $\pm$ 0.03).

Similarly, we found no significant main effects or interactions on mean listening count difference, p $>$ 0.05. When comparing the systems, $B1$ evaluators required 0.83 $\pm$ 0.56 more listens for the evaluation trials in comparison to the $B2$ evaluators (1.26 $\pm$ 0.51 listens). When completing the evaluation trials, non-native English speaking evaluators required 1.23 $\pm$ 0.45 more listens than native English speaking evaluators (0.86 $\pm$ 0.61 listens).

3.2. Correlation procedures with objective evaluations

For the evaluation trials, Pearson correlation procedures between PLDA scores to mean F_1, mean purity, and mean listening count metrics between systems and native and non-native English speaking evaluators Table 1 illustrates our findings.

4. Discussion

This study demonstrated that evaluators were able to use a clustering interface to link natural and anonymised speech record-

Table 1: *Pearson correlation results between PLDA scores and subjective results across anonymisation systems and native and non-native English speaking evaluators*

	System				Language			
	$B1$		$B2$		Native		Non-native	
Metric	ρ	p	ρ	p	ρ	p	ρ	p
F_1	0.05		0.7	***	0.46	*	0.4	**
Purity	0.36	*	0.7	***	0.46	*	0.42	**
Listening count	0.16		0.46	**	0.44	*	0.26	

where {*,**,***} mark significance for p <{0.05, 0.01, 0.001}

ings based on their perceived similarities. While we reported no significant differences between systems across the subjective metrics (F_1, purity, listening count), we observed a significant difference between native and non-native English speaking evaluators for mean F_1 difference between control and evaluation trials. We also reported that the subjective metrics correlated significantly to PLDA scores for the $B2$ system, but not the $B1$ system. Native and non-native English speaking evaluator performance correlated similarly and significantly, with the exception of listening count.

Our preliminary analysis revealed that evaluators who were presented $B1$-anonymised speech recordings required an additional 20 s (on average) to complete each trial, which suggests they found this task more difficult in comparison to the task of linking natural and $B2$-anonymised speech recordings. Although insignificant, we observed a decrease in performance when evaluators were tasked to link natural speech recordings with $B1$ anonymised speech recordings. On the other hand, non-native English speaking evaluators significantly lowered their accuracy when presented anonymised stimuli from either system. These findings suggest that the effectiveness of an anonymisation system can change depending on its users.

We reported no significant effects on cluster purity, which is a relatively novel metric. Our method of its calculation proposed that for each trial there was a maximum of four clusters each assigned to a different speaker. As 74 evaluators completed three trials, we expected to analyse 888 clusters, but instead observed a total of 852 total clusters, as 33 evaluators made less than four clusters per trial (53 total). This observation suggests our purity formula might require adaptations to better model the selection-making process of each individual evaluator. However, both F_1 and cluster purity correlated quite similarly correlated to the PLDA scores, which suggests it as a viable alternative to traditional binary metrics.

We hypothesised that the presence of anonymised speech would require evaluators to increase their number of listens, however, we reported no significant findings. The cluster task was designed to allow evaluators to personalise their engagements with the recordings, and thus employ a variety of listening strategies. Thus our observations suggest that the listening behaviours of evaluators did not change when they assessed anonymised speech recordings.

The significant correlations between subjective and objective results for $B2$ evaluation trials suggests evaluators clustered natural and anonymised speech recordings in a manner that was similar to the scores generated by our ASV model. The increase in the mean PLDA scores between natural and $B2$-anonymised speech recordings correlated to an increase in evaluator accuracy for both accuracy metrics (F_1, purity), as well as mean listening count per cluster. These observations were not consistent with the $B1$ system, which suggests that evaluator performance was unaffected and independent of the mean PLDA scores between speech recordings in a cluster. Regarding the general goals of voice privacy preservation, our findings suggest the $B1$ system is more effective at encumbering an attacker trying to access a private database. We reported similar significant correlations for both native and non-native English speaking evaluators between accuracy metrics and PLDA scores, which suggests spoken language familiarity did not affect performance when linking natural and anonymised speech recordings.

5. Conclusions

Our findings add to the growing number of studies focused on voice privacy and anonymisation, however, there remains many areas to develop. One critique of our study was the testing conditions, as evaluators were completely autonomous. While this setting mimics real-world environments, it was difficult to assess evaluator-specific factors ranging from general understanding of the task to technical issues, such as headphone use and internet connection. These factors might have influenced the performance of some evaluators, which underscore concerns regarding the current shift towards online perceptual-studies. Future work might consider comparing the performances of evaluators completing the clustering tasks both in- and out-side the laboratory setting.

Speaker recognition algorithms are designed to train and test on hours of speech recordings derived from hundreds of speakers. As we were limited to 33 and 39 evaluators for the $B1$ and $B2$ anonymisation systems, respectively, we were unable to examine any patterns associated with specific speakers. Our trial design allowed us to develop hundreds of tests without repeating the same combinations of speakers, utterances, and voice qualities. However, we might consider selecting a smaller set of male and female target and distractor speakers, so as to identify whether evaluators found the linking of natural and anonymised speech recordings belonging to particular speakers varied in difficulty. This selection process could be done by conducting preliminary speaker comparisons based on PLDA scores and assembling different speaker groups based on measured similarities.

We selected anonymised speech recordings from two anonymisation algorithms, however additional systems have been developed as part of the VoicePrivacy challenge. We aim to test these systems using similar methods to examine the limitations of linking natural and anonymisation speech recordings.

6. Acknowledgements

VoicePrivacy was born at the crossroads of projects VoicePersonae, COMPRISE (`https://www.compriseh2020.eu/`), and DEEP-PRIVACY. Project HARPOCRATES was designed specifically to support it. The authors acknowledge support by ANR, JST, and the European Union's Horizon 2020 Research and Innovation Program. They thank Teva Merlin for his development of the clustering interface.

7. References

[1] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado *et al.*, "Preserving privacy in speaker

and speech characterisation," *Computer Speech and Language*, vol. 58, pp. 441–480, 2019.

[2] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans *et al.*, "The VoicePrivacy 2020 Challenge evaluation plan," 2020. [Online]. Available: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf

[3] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech 2020*, 2020, pp. 1693–1697. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1333

[4] A. Nautsch, J. Patino, N. Tomashenko, J. Yamagishi, P.-G. Noé, J.-F. Bonastre, M. Todisco, and N. Evans, "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment," in *Proc. Interspeech 2020*, 2020, pp. 1698–1702. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1815

[5] P.-G. Noé, J.-F. Bonastre, D. Matrouf, N. Tomashenko, A. Nautsch, and N. Evans, "Speech Pseudonymisation Assessment Using Voice Similarity Matrices," in *Proc. Interspeech 2020*, 2020, pp. 1718–1722. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2720

[6] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, "A Comparative Study of Speech Anonymization Metrics," in *Proc. Interspeech 2020*, 2020, pp. 1708–1712. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2248

[7] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Speech Synthesis Workshop*, 2019, pp. 155–160.

[8] H. Hollien, R. Bahr, H. Kunzel, and P. Hollien, "Criteria for earwitness lineups," *International Journal of Speech Language and the Law*, vol. 2, pp. 143–153, 04 2013.

[9] B. O'Brien, A. Ghio, C. Fredouille, J.-F. Bonastre, and C. Meunier, "Discriminating speakers using perceptual clustering interface," in *Proc. XVII AISV Conference: Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications*, 2021.

[10] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/3443

[11] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," in *Interspeech*, 2020.

[12] S. McAdams, "Spectral fusion, spectral parsing and the formation of the auditory image," *Ph. D. Thesis, Stanford*, 1984.

[13] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," in *Proc. Interspeech*, 2021.

[14] U. Brandes, M. Gaertler, and D. Wagner, "Engineering graph clustering : Models and experimental evaluation," *First publ. in: ACM Journal of Experimental Algorithmics 12 (2007), Article 1.1*, vol. 12, 01 2007.

[15] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, 08 2015.

[16] 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)*. IEEE, 2018, pp. 5329–5333.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[18] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.