applied sciences

MDPI

Article

A Comparison of Hybrid and End-to-End ASR Systems for the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge

Juan M. Perero-Codosero [1,2,*], Fernando M. Espinoza-Cuadros [1,2,*] and Luis A. Hernández-Gómez [2,*]

[1] Sigma Technologies S.L.U., 28050 Madrid, Spain
[2] GAPS Signal Processing Applications Group, Universidad Politécnica de Madrid, 28040 Madrid, Spain
* Correspondence: jmperero@sigma-ai.com (J.M.P.-C.); fmespinoza@sigma-ai.com (F.M.E.-C.); luisalfonso.hernandez@upm.es (L.A.H.-G.)

Abstract: This paper describes a comparison between hybrid and end-to-end Automatic Speech Recognition (ASR) systems, which were evaluated on the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge. Deep Neural Networks (DNNs) are becoming the most promising technology for ASR at present. In the last few years, traditional hybrid models have been evaluated and compared to other end-to-end ASR systems in terms of accuracy and efficiency. We contribute two different approaches: a hybrid ASR system based on a DNN-HMM and two state-of-the-art end-to-end ASR systems, based on Lattice-Free Maximum Mutual Information (LF-MMI). To address the high difficulty in the speech-to-text transcription of recordings with different speaking styles and acoustic conditions from TV studios to live recordings, data augmentation and Domain Adversarial Training (DAT) techniques were studied. Multi-condition data augmentation applied to our hybrid DNN-HMM demonstrated WER improvements in noisy scenarios (about 10% relatively). In contrast, the results obtained using an end-to-end PyChain-based ASR system were far from our expectations. Nevertheless, we found that when including DAT techniques, a relative WER improvement of 2.87% was obtained as compared to the PyChain-based system.

Keywords: TV show speech-to-text transcription; ASR systems; hybrid DNN-HMM; end-to-end deep learning; domain adversarial training

Citation: Perero-Codosero, J.M.; Espinoza-Cuadros, F.M.; Hernández-Gómez, L.A. A Comparison of Hybrid and End-to-End ASR Systems for the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge. *Appl. Sci.* **2022**, *12*, 903. https://doi.org/10.3390/app12020903

Academic Editors: António Joaquim da Silva Teixeira, Francesc Alías, Valentin Cardeñoso-Payo, David Escudero-Mancebo and César González-Ferreras

Received: 22 December 2021
Accepted: 15 January 2022
Published: 17 January 2022

1. Introduction

Recently, the advancement of deep learning techniques has been able to improve the performance of Automatic Speech Recognition (ASR) systems. At the beginning, Deep Neural Networks (DNNs) became a fundamental part of conventional hybrid ASR systems [1]. According to some research studies [2], these models perform better in many scenarios with a small amount of training data, but they usually require strong context-dependent trees to train the models [3].

Nevertheless, end-to-end approaches are emerging [4,5] due to the reduction of the complexity associated with the training process. Whilst hybrid systems need to use Hidden Markov Model (HMM) state probabilities to train the outputs of a DNN, end-to-end systems are trained to map an input feature sequence to a sequence of characters [6,7]. Furthermore, the independence of intermediate modeling (e.g., acoustic, pronunciation, and language models) makes it easier to build an ASR model. They neither require any phoneme alignment for framewise cross-entropy, nor a sophisticated beam search decoder [8].

Several approaches have appeared such as Connectionist Temporal Classification (CTC) [4], the Recurrent Neural Network Transducer (RNN-T) [3], and the sequence-to-sequence attention-based encoder–decoder [5,9]. This trend presents an easy-to-use and easy-to-update pipeline. First, the training process does not have several stages, in which more than a single model would be involved. Second, the continuous advances in deep-learning-based technologies have allowed the quick development of powerful open-source libraries for machine learning, such as PyTorch [10] or TensorFlow [11], among others.

The promising results reported by many end-to-end ASR systems depend on the scenarios, as well as on the availability of datasets. Thus, end-to-end ASR models have achieved state-of-the-art results on the LibriSpeech database [12] and large public [13] or proprietary datasets [6]. These end-to-end models demand the availability of large training datasets [6], required for training very complex deep architectures [12]. However, in noisy scenarios and low-resource domains, such as CHiME-6 [14], end-to-end methods are still far from reaching the performance of HMM-based systems, as was reported, for example, in Ref. [15]. Thus far, end-to-end systems have not been able to overcome the best conventional hybrid models in those challenging conditions.

It is a fact that up to now, there is still a gap between end-to-end systems and hybrid models. To tackle this, some studies, such as SpeechStew [16], focused on demonstrating that a model is able to learn powerful transfer learning representations from a high volume of available speech data. The authors pointed out that training large models is expensive and not practical to perform frequently, but transfer learning allows fine-tuning a model pretrained on a combination of several public speech recognition datasets.

Recent developments focused on reducing this gap have reported good results, as was the case of ESPNet [17] or PyChain [18]. In PyChain, the end-to-end LF-MMI criterion, which is the state-of-the-art for hybrid models in Kaldi [19], is implemented by combining a single-stage training and a full parallelization under the PyTorch framework. Other speech recognition challenges, such as multichannel robust end-to-end ASR, have been addressed by a joint training of DNN-based front-end (speech enhancement) and back-end (speech recognition) models based on CTC-Attention and the RNN-T [15].

Besides that, to improve the performance of end-to-end ASR systems, a variety of techniques commonly applied in deep Learning have been introduced. Data augmentation techniques [20,21] have been developed to increase the quality and variety of training data following some criteria to improve the model robustness. Thus, a variety of scenarios can be simulated trying to cover the more challenging acoustic conditions in a cost-effective way.

Other works have been focused on the enhancement of deep acoustic models, where a sequence of local feature vectors is squeezed into a single global context vector [22], representing both speaker and environment information. In addition, model agnostic meta-learning has also been applied to rapidly adapt ASR models on cross-accented speech [23].

Other recent deep-learning-based techniques, such as Domain Adversarial Training (DAT) [24], have demonstrated that the model is able to reuse a latent space to improve performance on unseen input domains. Acoustic features must be robust to model the wide variety of speaker characteristics [25] and can play a relevant role in avoiding the bias in ASR systems with regard to diversity in gender, age, regional accents, and non-native accents, as was reported in Ref. [26]. To this end, DAT has been applied to ASR tasks by learning features invariant to different conditions, such as acoustic variabilities [27,28], accented speech [29], and inter-speaker feature variability [30].

In this paper, our aim was to contribute to the comparison of both hybrid and end-to-end ASR systems under the conditions of the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge [31]. This can be considered one of the aforementioned complex scenarios containing a variety of TV shows and broadcast news, in different noisy environments and challenging scenarios, such as TV debates. For this purpose:

- We firstly studied state-of-the-art techniques for hybrid and end-to-end ASR systems;
- We report the use of data augmentation techniques to improve our Kaldi-based hybrid ASR system presented in the IberSpeech-RTVE 2018 edition [32];
- Then, we evaluated a baseline end-to-end system on a real TV content dataset. We chose PyChain because it is based on the state-of-the-art LF-MMI approach, for which good results have been previously reported;
- Finally, looking to improve the end-to-end ASR system, we propose the use of DAT to learn features invariant to the environmental conditions and TV show format. Thus, we developed a novel improved version of the PyChain baseline including DAT. This

implementation allowed us to compare the performance of both end-to-end systems in the case of having low-computational or -speech data resources.

The rest of the paper is structured as follows. In Section 2, we describe the architecture of the ASR systems: a Kaldi-based hybrid ASR and two PyChain-based end-to-end systems. Section 3 explains the experimental protocols we followed under the IberSpeech-RTVE 2020 Challenge. Results are shown and discussed in Section 4. Section 5 presents related works comparing our systems with those submitted to IberSpeech-RTVE 2020 and other approaches in prior work. Finally, we present our conclusions in Section 6.

2. Architectures

2.1. DNN-HMM ASR

This system is based on the Sigma ASR system [32] submitted to the Albayzin-RTVE 2018 Speech-to-Text Challenge [33], where it was in the top 2 ranking for both closed- and open-condition evaluation.

This hybrid ASR system was built by using the Kaldi Toolkit [2]. The acoustic model is based on Deep Neural Networks and Hidden Markov Models (DNN-HMMs), following the so-called chain models [19], whose neural part is a subsampled Time-Delay Neural Network (TDNN) [34]. This implementation uses a 3-fold reduced frame rate at the output of the network.

We used the conventional feature pipeline that involves splicing 13-dimensional MFCC coefficients across 9 frames, followed by applying Linear Discriminant Analysis (LDA) to reduce the dimension to 40 and further decorrelation by means of Maximum Likelihood Linear Transform (MLLT). In addition, Feature-space Maximum Likelihood Linear Regression (fMLLR) was applied in a speaker-adaptive way. The input feature vectors were represented by 40-dimensional MFCC spliced coefficients across 7 frames and LDA+MLLT+fMLLR corresponding to 3 frames on each side of the central frame. In addition, 100-dimensional i-vectors were appended to the 40-dimensional acoustic space on each frame.

Our main conclusion from the results of Albayzin-RTVE 2018 [33] was the need for more robust DNN training, looking for accuracy improvements, required in the most challenging scenarios (street interviews, game shows, risky sports documentaries, etc.). The environmental robustness of acoustic models has been significantly improved by using multi-condition training data. However, the data collection process is very costly compared to the artificial generation of new training data, which has become a common alternative [20].

Thus, in our current contribution, we extended the amount of training data through data augmentation techniques. In particular, we added reverberation to the available training speech data following the approach presented in Ref. [35]. Depending on the expected scenarios and distances, different Room Impulse Responses (RIRs) can be used. They sample the room parameters and receiver position in the room and then randomly generate a number of RIRs according to different speaker positions. In short, three sets of simulated RIRs were applied: small room (1–10 m), medium room (10–30 m), and large room (30–50 m). The real computation was carried out at the feature extraction level, where the original data were mixed with their reverberated copies. The result was a 2-fold training set.

This data augmentation technique was added to other data augmentation techniques already used in the recipe followed in our previous system such as volume and speed perturbations [20,21].

2.2. End-to-End LF-MMI ASR

Aiming to explore new state-of-the-art end-to-end ASR systems, we evaluated an alternative to the developed Kaldi-based hybrid ASR system. That was the new end-to-end ASR Lattice-Free Maximum Mutual Information (LF-MMI) approach [19], which is also

used in Kaldi's chain models. Thus, we found it interesting to perform a reliable comparison of these two systems based on LF-MMI under IberSpeech-RTVE 2020 Challenge scenarios.

This end-to-end ASR system is based on PyChain [18], a powerful PyTorch-based implementation, which is intended to have an easy-to-use pipeline in which the data preparation and final decoding are carried out in Kaldi for efficiency, while data loading and network training are performed in PyTorch [10]. It should be noted that no alignment was necessary, i.e., the HMM-GMM training stage is not required, unlike other systems [36,37].

Data preparation consisted of both feature extraction (40-dimensional MFCC) and numerator/denominator graph (FSTs) generation. By following Kaldi's method for LF-MMI, HMM graphs were used for supervision. Consequently, the final LF-MMI loss function can be expressed as follows:

$$L_{MMI} = \sum_{u=1}^{U} \log \frac{P(X^{(u)}|\mathbb{G}_{num}^{(u)})}{P(X^{(u)}|\mathbb{G}_{den}^{(u)})} \quad (1)$$

where $X^{(u)}$ is the input frame sequences for the u-th utterance, while $\mathbb{G}_{num}$ and $\mathbb{G}_{den}$ are the numerator and denominator graph, respectively. These graphs are a combination of an n-gram phone Language Model (LM) with the acoustic part encoding all possible word sequences. As is widely known, $\mathbb{G}_{den}$ is generated from any possible transcription, while $\mathbb{G}_{num}$ makes use of the true transcription.

The probability distribution function (pdf) is used to estimate the likelihood of an HMM emission [2]. In this case, the network output and the occupation probability are computed from a pdf-index (pdf-id) instead of an HMM state. More specific details were presented in Ref. [18].

Once the data are loaded, the PyTorch model tries to simulate a TDNN [34] by including 1D dilated convolution in addition to batch normalization, ReLU, and dropout. This sequence is stacked in that order up to 6 layers with residual connections. At the end of the sequence, a fully connected layer is added (as described in Ref. [18]). From now on, this system is called the PyChain-based baseline system.

2.3. End-to-End LF-MMI ASR Applying Domain Adversarial Training

Different acoustic conditions of TV shows can have a negative impact on the PyChain-based baseline's performance. To reduce this effect, we explored the integration of DAT [24], trying to improve the PyChain-based baseline system. More specifically, in this approach, we aimed to make acoustic representations invariant to the domain of the TV show characteristics by using a Domain Adversarial Neural Network (DANN).

For this adversarial architecture, a training dataset denoted as $\{x_i, y_i, z_i\}_N^{i=1}$ is composed of x_i, which are the acoustic features, and y_i, z_i, which are the posteriors of the senones and the type of TV show, respectively.

Different from the PyChain-based baseline system training, in which the acoustic representation is trained so as to minimize the LF-MMI loss function, in DAT, the acoustic representations are learned adversarially against the secondary task (i.e., TV show classification). In this way, the domain-dependent information is suppressed in the representation, as it is irrelevant for the primary task (i.e., posterior classifier).

As can be seen in Figure 1, the parameters of our adversarial architecture consist of three parts, $\theta = \{\theta_x, \theta_y, \theta_z\}$, where θ_x denotes the parameters of the first layers of the TDNN used as the feature extractor and θ_y and θ_z denote the parameters of the pdf posteriors and the TV show classifier sub-networks, respectively.

Between the feature extractor and the TV show classifier, a Gradient Reversal Layer (GRL) [24] was implemented. In the forward propagation, the GRL keeps the input unchanged and reverses the gradient by multiplying it by a negative coefficient during the backpropagation.

According to [24], for this adversarial training, the objective function for the TV show classifier L_{dom} is defined as:

$$L_{dom}(\theta_x, \theta_z) = -\sum_{i=i}^{N} \log P(z_i|x_i; \theta_x, \theta_z) \quad (2)$$

The DNN acoustic model and the adversarial branch were jointly trained to optimize the following:

$$\min_{\theta_x, \theta_y} \max_{\theta_z} L_{MMI}(\theta_x, \theta_y) - \lambda L_{dom}(\theta_x, \theta_z), \quad (3)$$

where λ is a trade-off parameter between the pdf classification loss L_{MMI}, which corresponds to the LF-MMI loss defined in Equation (1), and the domain loss L_{dom}, related to the TV show classification task, which aims to make deep acoustic features invariant to the domain of the TV show characteristics.

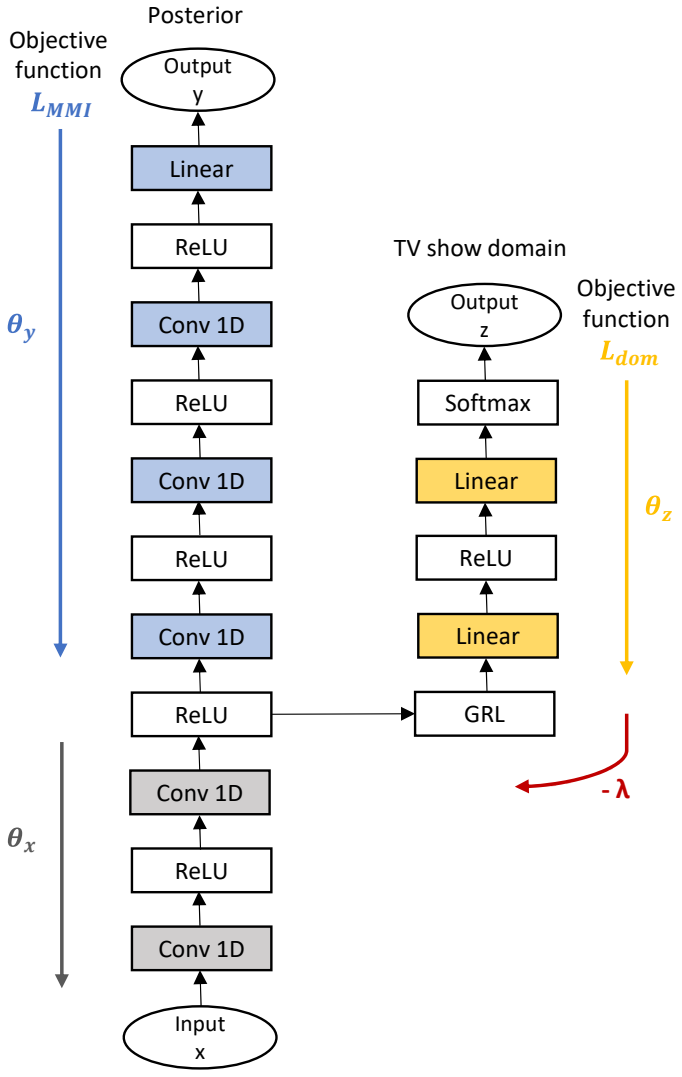


Figure 1. Architecture of the end-to-end LF-MMI approach applying DAT. An adversarial branch (TV show classifier) is added to the second layer of the main PyChain architecture (posterior classifier).

3. Experimental Setup

3.1. RTVE2020 Database

The proposed ASR systems were evaluated under the IberSpeech-RTVE 2020 Challenge conditions. The RTVE2020 Database [38] was provided to the participants. This is an extension of the RTVE2018 Database, which contains a collection of Spanish TV shows and

broadcast news from 2015 to 2019. The training partition consists of audio files, partially subtitled, presenting the following limitations:

- Subtitles were generated by means of a re-speaking procedure that sometimes changed the sentences and summarized what had been said, obtaining non-reliable transcriptions;
- Transcriptions were not supervised by humans. Only 109 h from the dev1, dev2, and test partitions contain human-revised transcriptions;
- Timestamps were not properly aligned with the speech signal.

Regarding all these limitations, we tried to avoid the use of these low-quality transcriptions, which could poorly model the acoustic space. As shown in Figure 2, we carried out a semi-supervised annotation process, which allowed training accurate acoustic models. First, a baseline acoustic model was trained using our own databases (see Section 3.2). Once the acoustic model was prepared, the unlabeled speech data were initially aligned to obtain a provisional transcription. To improve the quality of these automatic transcriptions, a human annotator team was responsible for the supervision process.

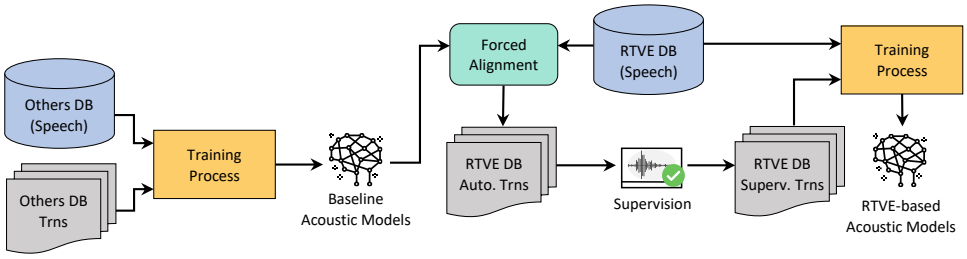


Figure 2. Entire process of obtaining high-quality transcriptions to train the acoustic models.

The RTVE training partition was prepared under this supervision process to obtain reliable transcriptions aligned with the speech signal. Hereafter, we were able to develop our first ASR models based on TV content [32].

Due to some limitations during the supervision process, two resulting datasets were used for training the systems: RTVE_train350 (350 h from RTVE training set) and RTVE_train100 (100 h from RTVE_train350). The validation datasets were 20% of the training data. It is worth noting that we tried to balance the partitions at any time, trying to cover the different scenarios represented in the whole RTVE dataset, such as political and economic news, in-depth interviews, debate and live magazines, among others. Consequently, for testing purposes, several datasets corresponding to 1 h in duration each were built from the RTVE_dev1 and RTVE_dev2 development partitions.

Finally, the RTVE2020 database was completed adding a collection of TV shows that belong to a wide range of genres and broadcasts from 2018 to 2019. This was composed of 70.3 h of human transcribed audio. It was used as the test partition for the Speech-to-Text Transcription Challenge (RTVE2020_test).

3.2. Other Databases

Additional datasets were added to train the system in an open training condition scenario:

The VESLIM database consists of 103 h of clean Spanish voice, where the speakers read a set of sentences. More details are in Ref. [39].

OWNMEDIA is composed of 162 h of TV shows, interviews, lectures, and several multimedia contents. It was used for training the baseline acoustic model, which allowed the initial alignment of the unlabeled speech data.

Finally, data augmentation techniques related to the hybrid ASR system were carried out by means of the reverberation database (http://www.openslr.org/28/, accessed on 14 January 2022), which was described in Section 2.1.

3.3. Training Setup

The acoustic model of the hybrid ASR system was trained using the RTVE_train350, VESLIM, and OWNMEDIA databases, following the SWBD Kaldi recipe for chain models. Some modifications were included following the ASpIRE recipe for multi-condition tasks.

Otherwise, end-to-end LF-MMI models without DAT were trained by using only RTVE_train100. This relatively small amount of data allowed a light training process to test the system performance. We used PyChain-example (https://github.com/YiwenShaoStephen/pychain_example, accessed on 14 January 2022) as a reference by adding some changes in terms of data loading and data parallelization to use more than one GPU.

The data preparation was carried out in Kaldi. To convert input features into PyTorch tensors, we used kaldi_io (https://github.com/vesis84/kaldi-io-for-python, accessed on 14 January 2022), as suggested in Ref. [18]. For the adversarial training, note that the number of pdf posteriors was $y_i = 62$, corresponding to the senones, and the number of TV shows was $z_i = 13$ because of the different TV shows that the RTVE_train100 partition contains. To reduce the bias effect due to unbalanced classes in TV shows at training time, the data were previously merged according to their acoustic characteristics. As a result, four new groups were defined (see also Table 1):

1. Live TV shows: a variety of content for the whole family;
2. Documentaries: show broadcasts about risky sports, adventure, street reports, and current information in different Spanish regions;
3. TV game shows: content related to comedy competitions, road safety, or culture dissemination, among others;
4. Interviews: moderated debates with analysis, political and economic news, and weather information.

Table 1. Description of the domain classes according to the number of samples and the characteristics of the TV shows. More details related to the TV shows are described in the RTVE2020 Database specifications [38].

Class	# of Samples	Examples of TV Shows
1. live TV shows	11,239	La Mañana
2. documentaries	4671	Al filo de lo Imposible, Comando Actualidad, España en Comunidad
3. TV-game shows	7995	Arranca en Verde, Dicho y Hecho, Saber y Ganar
4. interviews	22,194	Latinoamerica 24H, La Tarde en 24H, Millenium

As a consequence, the labels of the training data for the adversarial branch (i.e., the TV show classifier sub-network) are defined as $z_i = 0, 1, 2, 3$. In the adversarial architecture, the second hidden layer of the TDNN was used as the input to the adversarial branch, which consisted of a dense layer of size 384 and the ReLU activation function, followed by a softmax output layer, whose output dimension corresponded to the number of TV shows (i.e., 4). The cross-entropy loss function was used in the adversarial training. To select the optimal trade-off parameter λ, several values were tested. The best performance was achieved for $\lambda = 0.041$. In addition, all the systems were evaluated with the same 3-gram LM. As described in Ref. [32], it was trained on several corpora: subtitles provided in the RTVE2018 Database, supervised transcriptions, news between 2015 and 2018, interviews, and file captions.

3.4. Resources

Several computational resources were required to carry out this work. A server with 2 Xeon E5-2630V4, 2.2 GHz, 10C/20 TH, and 3 GPUs Nvidia GTX 1080 Ti was used for the hybrid ASR system. GPU calculation was necessary for the DNN stage, and only CPU mode was used for the HMM stage and final decoding.

4. Results

4.1. Hybrid ASR

The proposal of applying data augmentation techniques to improve the hybrid ASR performance was fulfilled. The addition of reverberation to our whole training dataset (over 600 h of speech) improved the performance in most of the scenarios represented by every TV show set. As shown in Table 2, the model applied to the Comando Actualidad (CA) dataset achieved a relative improvement of around 10%, as compared to the baseline system. This might be due to the trained model having learned to model these speech artifacts that can appear in the challenging scenarios described in Section 2.1. However, the improvements in the rest of the TV shows were not so remarkable (e.g., 20H or LM) because the contents were related to daily news with more favorable acoustic conditions.

As we mentioned in Ref. [32], the reference master of the transcriptions was not reviewed. As usual, we evaluated the possible impact of transcription errors by means of a new test using an external dataset. It consisted of 3.5 h of TV news broadcasts (similar to 20H). Applying the reverb-trained models of our Kaldi-based hybrid system, we reduced the WER from 8.51% to 7.96%, being our new best results achieved so far. Table 3 shows that data augmentation also maintained the WER improvement of around 10% relative on the RTVE2020 test partition.

Table 2. WER (%) on the different datasets for hybrid and end-to-end ASR systems. Each one of the evaluation datasets contains one hour of speech. In bold, the improvements of the Kaldi-based system after applying reverberated data augmentation, and the improvements related to the Pychain-based system after applying DAT.

	20H_dev1	AP_dev1	CA_dev1	LM_dev1	Mill_dev1	LN24H_dev1
Hybrid ASR						
Kaldi-based baseline [32]	14.88	20.94	49.55	21.44	17.01	24.13
Reverb. data augmentation	**14.76**	21.00	**44.69**	**21.03**	**16.42**	**23.62**
Kaldi-based baseline (RTVE_train100)	16.09	22.32	51.23	23.02	17.70	25.53
End-to-end LF-MMI ASR						
PyChain-based baseline	23.66	33.31	59.34	29.95	35.09	25.08
Domain adversarial training	**23.53**	**32.99**	**59.25**	**29.91**	**34.67**	25.16

Table 3. WER (%) on the RTVE2020 test partition for all the systems. Results were obtained after the submission.

	RTVE2020_test
Hybrid ASR	
Kaldi-based baseline [32]	31.01
Reverb. data augmentation	27.68
End-to-end LF-MMI ASR	
PyChain-based baseline	40.90
Domain adversarial training	42.89

4.2. End-to-End LF-MMI ASR

Our PyChain-based baseline had a good performance in relation to the number of parameters and the easier training process compared to other end-to-end frameworks. The WER achieved for standard TV news (e.g., 20H, LN24H) was between 23% and 26%, as Table 2 shows. These results are within the expected range where commercial ASR systems operate.

To compare both hybrid and end-to-end systems, we also trained a hybrid model by using only the RTVE_train100 partition. In this case, multi-condition data augmentation was not applied. The PyChain-based system was still far from the Kaldi-based hybrid system, with a WER increase of 17% in the worst-case scenario. This gap could be reduced with the application of some data augmentation techniques (e.g., speed or volume pertur-

bation). Despite the fact that augmented data caused a slight improvement of the WER for the PyChain system [18], TV shows can contain some acoustic characteristics that could be better modeled by some audio perturbations.

On the other hand, we evaluated the effect of learning acoustic representations invariant to the TV show domain. After applying DAT, the results in Table 2 showed improvements, in terms of the WER, up to 2.87% as compared to the PyChain-based baseline. We are aware that those results are still far from the Kaldi-based hybrid ASR system based on the DNN-HMM. Nevertheless, the results in Table 2 gave us an insight into how the use of DAT in the end-to-end LF-MMI model can improve its performance in most of the scenarios. Furthermore, it seemed that DAT was able to generate deep acoustic features invariant to different TV shows with different acoustic conditions without the need for data augmentation techniques.

In addition, Table 3 shows that applying DAT did not reduce the WER on the RTVE2020 test partition. The main reason was DAT alleviated the labeled domain conditions in the training dataset. Thus, invariant features were trained without regarding these unseen external factors.

Finally, regarding computational requirements for speech transcription, we considered that PyChain carries out two main stages: a first decoding stage and a second four-gram rescoring stage. Real-Time factors (RT) for different test partitions are presented in Table 4. In all cases, the two GPU resources described in Section 3.4 were used. The results in Table 4 show that time requirements depend on the characteristics of the test partitions. Less time is required to transcribe the best acoustic conditions and less challenging scenarios. This makes sense as far as the confusion of the graph model is less complex to transcribe accurately.

Table 4. Real-Time factor (RT) for the different stages carried out in the PyChain-based baseline according to the different datasets.

Datasets	Decoding	LM Rescoring
20H_dev1	0.033	0.115
AP_dev1	0.035	0.175
CA_dev1	0.225	1.976
LM_dev1	0.092	0.450
Mill_dev1	0.082	0.442
LN24H_dev1	0.383	0.148

5. Related Works

As the evaluation of our systems was carried out under the conditions of the IberSpeech-RTVE 2020 Challenge [31], we can now compare our work to other ASR systems participating in this challenge and thus also trained on this specific domain related to TV programs. As a first general comment, we can say that all the results for the developed ASR systems showed that end-to-end systems are still far from hybrid systems in the challenging conditions of RTVE 2020. Kocour et al. [40] developed an end-to-end system based on the wav2letter architecture [7], which was not able to generalize very well on the acoustic conditions of the RTVE2020 database, reporting a WER of at least 13% higher than their best hybrid ASR system. Álvarez et al. [41] presented a Quartznet-based [42] ASR implementation showing promising results due to the use of more than a hundred hours of speech data. Nevertheless, the results in terms of the WER from that system were 9% worse when compared to other hybrid systems they developed for the challenge.

Furthermore, the acoustic conditions of databases provided in other international challenges, such as CHiME-6 [14], have also been responsible for the poor performance of end-to-end ASR systems. Different approaches based on RNNs and transformers, along with the RNN-T and CTC-Attention [15] have been evaluated on one of the CHiME partitions, concluding that speech-enhancement techniques contribute to the reduction of to the gap between end-to-end and hybrid systems. The difficulty of obtaining high-

quality labeled speech leads to emergent machine-learning-based paradigms, such as Self-Supervised Learning (SSL), whose goal for ASR is to learn powerful speech representations from unlabeled examples (e.g., wav2vec 2.0 [43]). Other recent works have tried to mitigate the effect of limited training data [44] or noisy environment conditions [45]. In the case of [44], the use of CTC and end-to-end LF-MMI to fine-tune a wav2vec 2.0 model showed similar performance even for out-of-domain and cross-lingual adaptation. Regarding [45], the authors integrated SSL with contrastive learning from original–noisy speech pairs to model representations with noise robustness.

Along the same lines, we proposed the application of DAT in our baseline end-to-end ASR systems as an alternative to obtain robust features invariant to the domain, i.e., different acoustic conditions of TV shows. It is a fact that DAT is beneficial for building robust embeddings. In recent works, it has been even integrated with wav2vec embeddings to have an accent-robust speech recognition [46]. The authors reported good results when no accent labels were available for training.

6. Conclusions and Future Work

In this paper, we developed both hybrid and end-to-end ASR approaches exploring some techniques to improve the performance of Speech-to-Text tasks under IberSpeech-RTVE 2020 Challenge. We showed that Hybrid DNN-HMMs can be adapted to the TV show domain by means of multi-condition data augmentation. The addition of reverberated data to the training data decreased the WER significantly (10% relative). A WER of 7.96% was achieved in better conditions. We demonstrated that the lack of data augmentation techniques could be the main reason of the gap between the Kaldi-based hybrid system and PyChain-based system. Moreover, a higher volume of data used to train the end-to-end system could contribute to increasing the ASR performance. However, other easy-to-apply techniques, such as DAT, can overcome this gap, yielding improvements in end-to-end ASR systems. We found that using DAT, acoustic features invariant to different TV domains can be learned, achieving a WER improvement of 2.87%.

The IberSpeech-RTVE 2020 Challenge has provided some findings pointing out that end-to-end approaches are close to being competitive (between 28% and 40% WER) by using more than 600 h of speech. However, they are still far from the hybrid models.

As future work, besides data augmentation, we believe that exploring speech-enhancement techniques could help to close the performance gap between hybrid and end-to-end systems. In addition, unsupervised machine learning methods (e.g., clustering) or automatic perceptual speech quality methods (e.g., PESQ) could contribute to a more accurate TV show classification prior to DAT. Finally, we also believe that the combination of DAT with a self-supervised approach could be useful to achieve significant improvements in ASR systems through robust embeddings trained even without supervised data.

Author Contributions: Conceptualization, J.M.P.-C., F.M.E.-C. and L.A.H.-G.; data curation, J.M.P.-C.; formal analysis, J.M.P.-C., F.M.E.-C. and L.A.H.-G.; investigation, J.M.P.-C., F.M.E.-C. and L.A.H.-G.; methodology, J.M.P.-C.; software, J.M.P.-C. and F.M.E.-C.; supervision, F.M.E.-C. and L.A.H.-G.; validation, J.M.P.-C. and F.M.E.-C.; writing–original draft, J.M.P.-C.; writing–review & editing, F.M.E.-C. and L.A.H.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: RTVE 2020 database was used under a license agreement. It is available upon request in http://catedrartve.unizar.es/rtvedatabase.html, accessed on 14 January 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
2. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
3. Graves, A. Sequence transduction with recurrent neural networks. *arXiv* **2012**, arXiv:1211.3711.
4. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
5. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4960–4964.
6. Chiu, C.C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E.; et al. State-of-the-art speech recognition with sequence-to-sequence models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778.
7. Collobert, R.; Puhrsch, C.; Synnaeve, G. Wav2letter: An end-to-end convnet-based speech recognition system. *arXiv* **2016**, arXiv:1609.03193.
8. Zeyer, A.; Irie, K.; Schlüter, R.; Ney, H. Improved training of end-to-end attention models for speech recognition. *arXiv* **2018**, arXiv:1805.03294.
9. Bahdanau, D.; Chorowski, J.; Serdyuk, D.; Brakel, P.; Bengio, Y. End-to-end attention-based large vocabulary speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4945–4949.
10. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 8026–8037.
11. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning. In Proceedings of the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
12. Zhang, Y.; Qin, J.; Park, D.S.; Han, W.; Chiu, C.C.; Pang, R.; Le, Q.V.; Wu, Y. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv* **2020**, arXiv:2010.10504.
13. Chen, G.; Chai, S.; Wang, G.; Du, J.; Zhang, W.Q.; Weng, C.; Su, D.; Povey, D.; Trmal, J.; Zhang, J.; et al. GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 h of Transcribed Audio. *arXiv* **2021**, arXiv:2106.06909.
14. Watanabe, S.; Mandel, M.; Barker, J.; Vincent, E.; Arora, A.; Chang, X.; Khudanpur, S.; Manohar, V.; Povey, D.; Raj, D.; et al. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv* **2020**, arXiv:2004.09249.
15. Andrusenko, A.; Laptev, A.; Medennikov, I. Towards a competitive end-to-end speech recognition for chime-6 dinner party transcription. *arXiv* **2020**, arXiv:2004.10799.
16. Chan, W.; Park, D.; Lee, C.; Zhang, Y.; Le, Q.; Norouzi, M. SpeechStew: Simply mix all available speech recognition data to train one large neural network. *arXiv* **2021**, arXiv:2104.02133.
17. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N.E.Y.; Heymann, J.; Wiesner, M.; Chen, N.; et al. Espnet: End-to-end speech processing toolkit. *arXiv* **2018**, arXiv:1804.00015.
18. Shao, Y.; Wang, Y.; Povey, D.; Khudanpur, S. PyChain: A Fully Parallelized PyTorch Implementation of LF-MMI for End-to-End ASR. *arXiv* **2020**, arXiv:2005.09824.
19. Povey, D.; Peddinti, V.; Galvez, D.; Ghahremani, P.; Manohar, V.; Na, X.; Wang, Y.; Khudanpur, S. *Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI*; Interspeech: San Francisco, CA, USA, 2016; pp. 2751–2755.
20. Peddinti, V.; Chen, G.; Manohar, V.; Ko, T.; Povey, D.; Khudanpur, S. *JHU ASpIRE System: Robust LVCSR with TDNNS, iVector Adaptation and RNN-LMS*; In Proceedings of the IEEE 2015 Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 539–546.
21. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
22. Han, W.; Zhang, Z.; Zhang, Y.; Yu, J.; Chiu, C.C.; Qin, J.; Gulati, A.; Pang, R.; Wu, Y. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv* **2020**, arXiv:2005.03191.
23. Winata, G.I.; Cahyawijaya, S.; Liu, Z.; Lin, Z.; Madotto, A.; Xu, P.; Fung, P. Learning Fast Adaptation on Cross-Accented Speech Recognition. *arXiv* **2020**, arXiv:eess.AS/2003.01901.
24. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
25. Dhakal, P.; Damacharla, P.; Javaid, A.Y.; Devabhaktuni, V. A near real-time automatic speaker recognition architecture for voice-based user interface. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 504–520. [CrossRef]

26. Feng, S.; Kudina, O.; Halpern, B.M.; Scharenborg, O. Quantifying bias in automatic speech recognition. *arXiv* **2021**, arXiv:2103.15122.
27. Serdyuk, D.; Audhkhasi, K.; Brakel, P.; Ramabhadran, B.; Thomas, S.; Bengio, Y. Invariant representations for noisy speech recognition. *arXiv* **2016**, arXiv:1612.01928.
28. Shinohara, Y. *Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition*; Interspeech: San Francisco, CA, USA, 2016; pp. 2369–2372.
29. Sun, S.; Yeh, C.F.; Hwang, M.Y.; Ostendorf, M.; Xie, L. Domain adversarial training for accented speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4854–4858.
30. Meng, Z.; Li, J.; Chen, Z.; Zhao, Y.; Mazalov, V.; Gang, Y.; Juang, B.H. Speaker-invariant training via adversarial learning. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5969–5973.
31. Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; de Prada, A. Albayzin Evaluation: IberSPEECH-RTVE 2020 Speech to Text Transcription Challenge. 2020. Available online: http://catedrartve.unizar.es/reto2020/EvalPlan-S2T-2020-v1.pdf (accessed on 14 January 2022).
32. Perero-Codosero, J.M.; Antón-Martín, J.; Merino, D.T.; Gonzalo, E.L.; Gómez, L.A.H. *Exploring Open-Source Deep Learning ASR for Speech-to-Text TV Program Transcription*; IberSPEECH: Valladolid, Spain, 2018; pp. 262–266.
33. Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; de Prada, A. Albayzin 2018 evaluation: The iberspeech-RTVE challenge on speech technologies for spanish broadcast media. *Appl. Sci.* **2019**, *9*, 5412. [CrossRef]
34. Peddinti, V.; Povey, D.; Khudanpur, S. A time delay neural network architecture for efficient modeling of long temporal contexts. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
35. Ko, T.; Peddinti, V.; Povey, D.; Seltzer, M.L.; Khudanpur, S. A study on data augmentation of reverberant speech for robust speech recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5220–5224.
36. Ravanelli, M.; Parcollet, T.; Bengio, Y. The pytorch-kaldi speech recognition toolkit. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6465–6469.
37. Can, D.; Martinez, V.R.; Papadopoulos, P.; Narayanan, S.S. Pykaldi: A python wrapper for kaldi. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5889–5893.
38. Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; de Prada, A. RTVE2020 Database Description. 2020. Available online: http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf (accessed on 14 January 2022) .
39. Toledano, D.T.; Gómez, L.A.H.; Grande, L.V. Automatic phonetic segmentation. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 617–625. [CrossRef]
40. Kocour, M.; Cámbara, G.; Luque, J.; Bonet, D.; Farrús, M.; Karafiát, M.; Veselỳ, K.; Ĉernockỳ, J. BCN2BRNO: ASR System Fusion for Albayzin 2020 Speech to Text Challenge. *arXiv* **2021**, arXiv:2101.12729.
41. Alvarez, A.; Arzelus, H.; Torre, I.G.; González-Docasal, A. *The Vicomtech Speech Transcription Systems for the Albayzın-RTVE 2020 Speech to Text Transcription Challenge*; IberSPEECH: Virtual Valladolid, Spain, 2021; pp. 104–107.
42. Kriman, S.; Beliaev, S.; Ginsburg, B.; Huang, J.; Kuchaiev, O.; Lavrukhin, V.; Leary, R.; Li, J.; Zhang, Y. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Barcelona, Spain, 4–8 May 2020; pp. 6124–6128.
43. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv* **2020**, arXiv:2006.11477.
44. Vyas, A.; Madikeri, S.; Bourlard, H. Comparing CTC and LFMMI for out-of-domain adaptation of wav2vec 2.0 acoustic model. *arXiv* **2021**, arXiv:2104.02558.
45. Wang, Y.; Li, J.; Wang, H.; Qian, Y.; Wang, C.; Wu, Y. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. *arXiv* **2021**, arXiv:2110.04934.
46. Li, J.; Manohar, V.; Chitkara, P.; Tjandra, A.; Picheny, M.; Zhang, F.; Zhang, X.; Saraf, Y. Accent-Robust Automatic Speech Recognition Using Supervised and Unsupervised Wav2vec Embeddings. *arXiv* **2021**, arXiv:2110.03520.

Chapter 4

Discussion and Future Work

This section presents a summary and discussion of the main results obtained as part of this Thesis. Since all the implementation details and exhaustive results are included in the publications of the compendium, we present a global perspective of the results regarding the main and specific objectives defined as the starting point of the Thesis.

4.1 Results

In this Thesis we have researched adversarial learning, within Deep Neural Networks, in three speech applications. The main goal was to obtain results that can be used to evaluate the benefits of using adversarial learning to derive speech features invariant to undesired sources of variability.

Next, we briefly summarize the main results for the specific objectives (O) outlined and defined in Section 1.2:

O1. Study of adversarial learning techniques.

The review of the state-of-the-art adversarial learning techniques was successfully addressed. Adversarial learning fundamentals were studied to properly understand the method and the evolution of this technique as a type of adaptation towards a supervised method. As a result, an exhaustive search of DAT-related approaches in prior art was obtained. Relevant related works have been listed in Table 2.1 where we indicate the primary and secondary tasks, i.e., source and domain respectively, together with the improvement reported by each reference.

O2. Application of the adversarial learning techniques for the Assessment of Obstructive Sleep Apnea from speech.

First, we made a comparison between two paradigm models (i-vectors or x-vectors embeddings) to determine which one accurately estimates the severity of OSA voices (e.g., AHI) and other

clinical variables (e.g., AGE). Results obtained when predicting AHI from i-vectors and x-vectors show that AHI predictions from x-vectors are poorer than the ones using i-vectors: MAE=13.45, $\rho = 0.32$ (i-vectors); MAE=14.13, $\rho = 0.18$ (x-vectors). This initial study also showed the weaker impact of OSA on speech compared to the effect of speaker's age. It can make reasonable trying to model and differentiate among extreme OSA levels.

Consequently, a second study was addressed to classify OSA extreme cases (AHI$\leq$ 10 and AHI$\geq$ 30). We focused on those AHI ranges because better and more accurate results have been previously reported in (Solé-Casals et al. (2014); Espinoza-Cuadros et al. (2016); **?**). We implemented a DAT approach to derive speech features related to OSA but invariant to AGE and body mass index (BMI). AGE was considered because of the demonstrated correlation between AGE and speech, which is much stronger than between AHI and speech. BMI was also studied as it is the clinical variable with the strongest correlation with AHI.

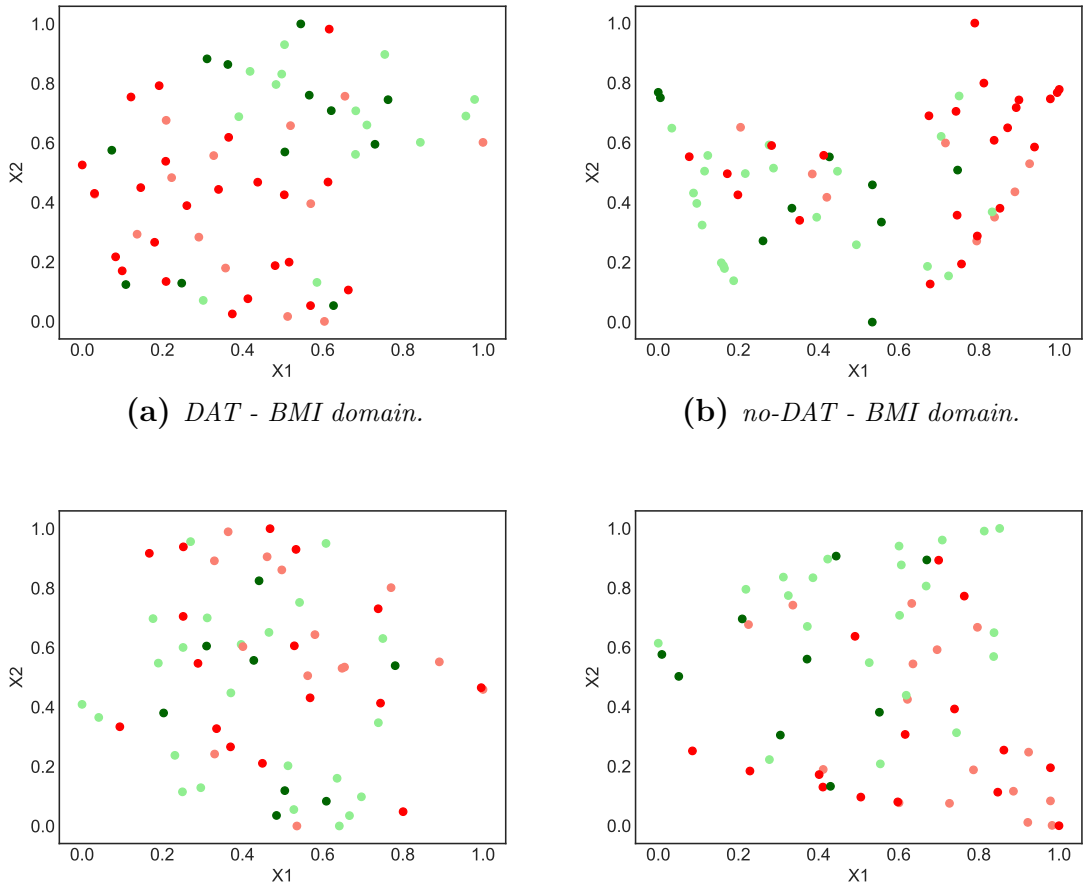
Our results show a significant increase in classification accuracy (from 69.39% to 76.60%) when DAT was applied assuming BMI as the adversarial domain. On the other hand, classification accuracy decreases when the AGE is considered as the adversarial domain.

While trying to understand and illustrate our results, we use t-SNE projection Maaten and Hinton (2008) to visualize the distribution of the speech feature embeddings before and after applying DAT. It should be noted that no-DAT visualization corresponds to i-vectors as inputs to the proposed DANN architecture. DAT projections represent domain invariant (AGE or BMI in each case) features generated by the DANN feature generator $G_f(x, \theta_f)$ (see Figure 4.1). As can be observed in these figures, the feature distribution after domain adaptation for BMI produces clearly the different clusters corresponding to extreme OSA cases. While for AGE the application of DAT results in separability between OSA classes not seeming to improve.

O3. Application of the adversarial learning techniques for Speech Anonymization.

For this objective, adversarial learning has been evaluated following the privacy preservation scenario proposed in the VoicePrivacy Initiative Tomashenko et al. (2020b) and using a baseline anonymization framework consisting of three steps: 1) Feature extraction, 2) x-vector anonymization, and 3) speech synthesis. An Autoencoder-Adversarial Network (AAN) was proposed to reconstruct the input x-vector, with adversarial branches to mitigate speaker characteristics as speaker identity, gender, or accent.

Our initial results were related to studying the behavior of the trade-off hyperparameter (λ) when training the AAN. This hyperparameter controls the relationship between the primary task, i.e., reconstructing the speech feature vector, and the secondary task, i.e., classifying the



(a) *DAT - BMI domain.* **(b)** *no-DAT - BMI domain.*

(c) *DAT - Age domain.* **(d)** *no-DAT - Age domain.*

Figure 4.1: *The effect of adaptation on the distribution of the extracted features. The figure show t-SNE Maaten and Hinton (2008) visualizations of the DNN's activations: (a) and (c) where adaptation procedure was incorporated into training, and (b) and (d) where no adaptation was performed. Dark and light colors distinguish the domain class. Green (AHI≤10) and red (AHI≥30) colors distinguish the OSA extreme-cases class*

speaker characteristics which we would like to remove. During the AAN training, different λ values, ranging from $\lambda = 0$ to 12, during backpropagation through a GRL were evaluated to find the optimal value that helps the AAN to perform the best in the anonymization framework. As the selection criteria, we aim to find the best trade-off between achieving the greatest anonymization and preserving the speech content intelligibility as much as possible.

For this purpose, during the first step, the standalone AAN model was evaluated using the test partition from the VoxCeleb-1 dataset, as described in sub-Section 2.3.2. In Figure 4.2, we can observe that the degradation in the classification performance in the speaker adversarial branch, in terms of accuracy, is drastically reduced at below 1%. This effect is also shown in the gender and accent branches, where the degradation is less than that of the speaker branch.

In this same vein, this effect is also presented in the distance between the reconstructed

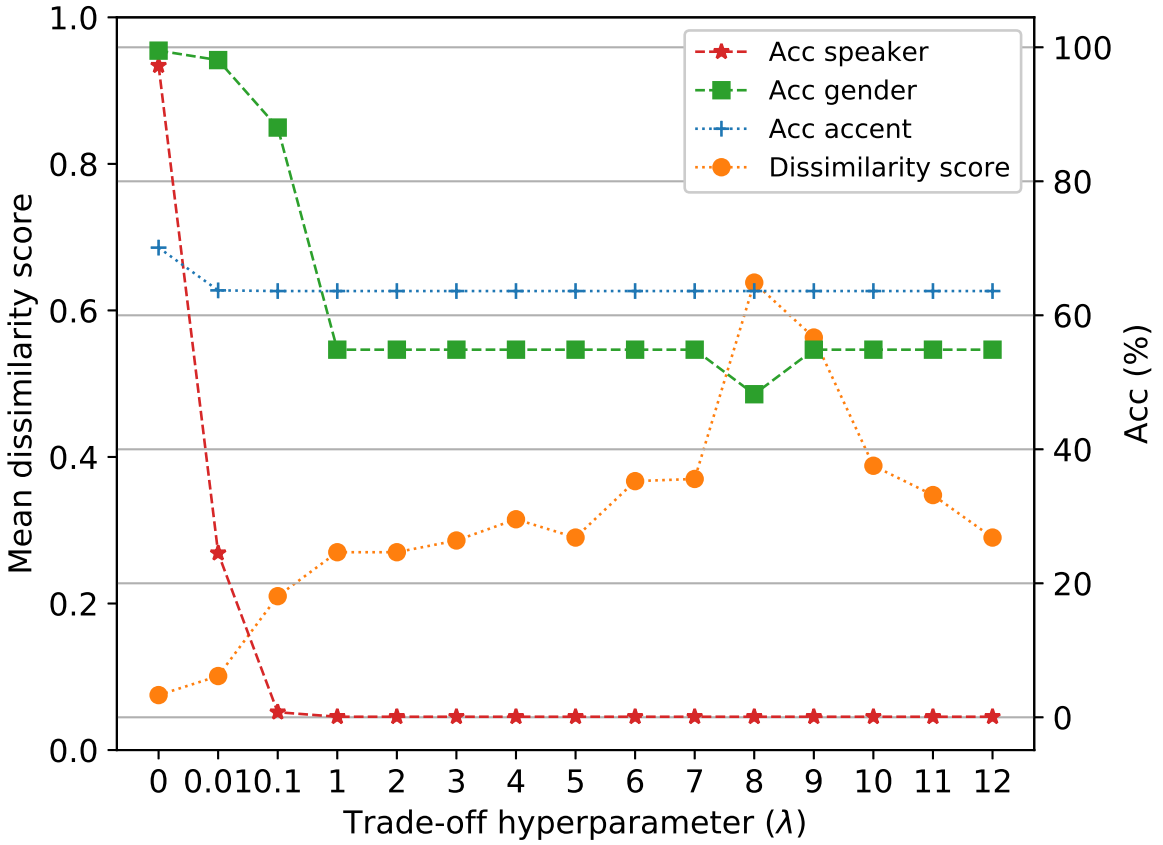


Figure 4.2: *AAN standalone evaluation using a test partition from VoxCeleb-1 dataset. Results are shown in terms of dissimilarity score and classification rate (ACC) for each adversarial branch (i.e., speaker, gender and accent).*

and original x-vectors. This distance is measured in terms of the dissimilarity score, defined as $1 - cos(x1, x2)$ where the higher λ the higher dissimilarity score, as shown in Figure 4.2. This increase in the dissimilarity score and the decrease in performance in gender classification end at $\lambda = 8$, whereas for $\lambda > 8$, the dissimilarity score starts to decrease. Thus far, the test results indicate that the best performance for removing speaker information from the x-vector is achieved using $\lambda = 8$.

In the second step, to evaluate the anonymization performance, the range of hyperparameter values was evaluated in the baseline framework using the VCTK-dev different subset. The original enrollment-anonymized trial attack scenario was used for this evaluation. As shown in Figure 4.3, similar to the AAN standalone evaluation, the higher the hyperparameter λ, the greater the EER and dissimilarity for $\lambda \leq 8$, whereas the WER is relatively stable across all hyperparameter values. This tendency is not usual in other x-vector anonymization-based systems where the higher the distance, the higher the degradation in the WER.

Despite of remaining the utility, i.e., low WER values meaning intelligibility preservation, at

$\lambda > 8$, the anonymization performance decreases in terms of EER and dissimilarity score. This effect may be caused by trying to force adversarial branches to remove the speaker information as much as possible. There is a lambda value, which is the limit in which no more degradation is produced in adversarial branches. Thus, the best trade-off between anonymization and speech content intelligibility preservation was $\lambda = 8$, where new generated synthetic voices are able to hide the speaker identity and characteristics while maintaining the speech content.

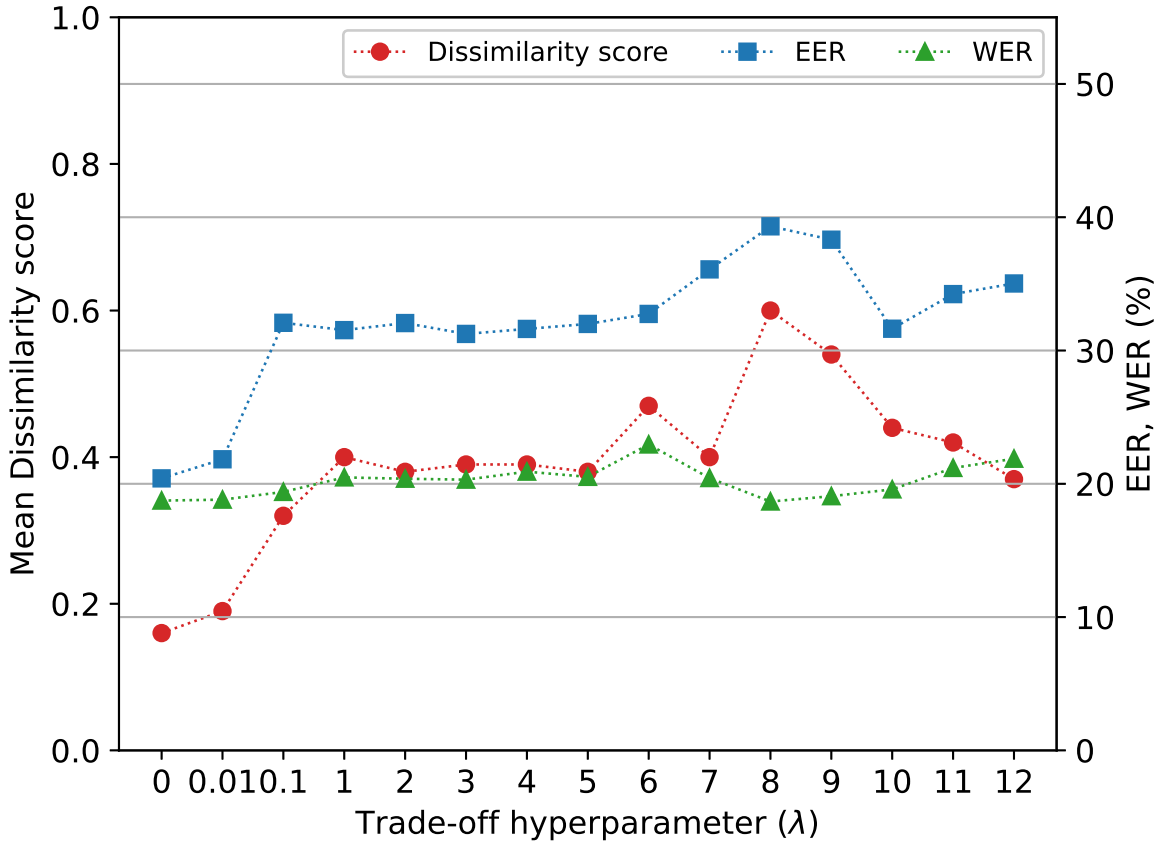


Figure 4.3: *EER(%), WER(%) and Mean dissimilarity score values obtained by out x-vector anonymization approach for different trade-off hyperparameters (λ) values for original enrollment-anonymized trial attack scenario evaluated on VCTK-dev different datset.*

The second experiment has the objective of evaluating the anonymization performance of our AAN approach. The evaluation was carried out under the VoicePrivacy 2020 Challenge framework, to ensure a transparent solution for speaker anonymization tested on standard datasets and in contrast with other benchmark solutions.

The ASV evaluation was carried out under different attack scenarios already defined in sub-section 2.3.3. These scenarios assess the effectiveness of anonymization in certain contexts; however, under the assumption that we aim to hide the speaker identity as much as possible while preserving the intelligibility of the spoken content, our goal is to achieve a high anonymization

performance, because the higher the EER, the better the anonymization.

We compared the results achieved using our x-vector anonymization approach for the best trade-off hyperparameter selected, i.e., $\lambda = 8$, to those of the baseline. In this case, our x-vector anonymization approach achieves a higher anonymization performance in the (a-enroll, a-trial) scenario when compared to the baseline, achieving improvements of up to 9% in terms of EER. This improvement in privacy also allows the speech content intelligibility to be maintained in terms of WER. This result is obtained by using multiple adversarial domains (speaker identity, gender and accent) instead of using one of them as a single adversarial domain.

Furthermore, for all attack scenarios and for both female and male populations, the EER was above 10%, and higher values were generally observed through multiple domain anonymization. The WER results also show a better performance for the multiple-domain case.

This is a reasonable result, because better privacy is expected when multiple adversarial domains are used. Nevertheless, anonymization results for individual domains, such as gender and accent, also show a reasonable performance. This demonstrates the capability of the proposed system to be applied under different scenarios, where only the gender or accent information of the speaker needs to be protected. In addition, the limited performance of the EER for the accent domain can be observed. This may be due to the limitation of VoxCeleb-1 dataset, where the accent classes are highly unbalanced.

Finally, we compared our x-vector anonymization approach to other systems in the literature, i.e., x-vector anonymization methods submitted to the VoicePrivacy Challenge. Our method achieved a high performance in terms of privacy, in other words the EER is one of the highest values in preserving the intelligibility, even improving over the baseline result; meanwhile WER result is practically the same as the Baseline.

O4. Application of the adversarial learning techniques for Automatic Speech Recognition.

To address this objective, we firstly developed and evaluated an end-to-end ASR system which was able to be compared to other hybrid systems, such as Kaldi-based approaches. More specifically, we chose PyChain, which is an end-to-end approach providing a good ASR performance in relation to the number of parameters and its easier training process. The WER achieved for standard TV news was between 23% and 26%. These results are within the expected range in which commercial ASR system operate. To easily compare both hybrid and end-to-end systems, we also trained a hybrid model using only the RTVE_train100 partition. PyChain-based system was still far from the Kaldi-based hybrid system, with a WER increase of 17% in the worst-case scenario. This gap could be reduced with the application of some data augmentation techniques.

Despite the fact that augmented data caused a slight improvement of the WER for the Pychain system (Shao et al. (2020)), TV shows can contain some acoustic characteristics that could be better modeled by some audio perturbations.

From this starting point, we implemented adversarial learning, DAT, within the end-to-end architecture, which simulates a TDNN as Kaldi chain models. Thus, we were able to evaluate the effect of learning acoustic representations invariant to the TV show domain. After applying DAT, the results on RTVE2018 speech data (the same types of TV shows as those used in the training stage) showed a small improvement, in terms of WER, up to 2.87% as compared to the PyChain-based baseline.

Another evaluation was accomplished taking the blind RTVE2020 test partition. At the end of the challenge, the organization provided the ground-truth labels, i.e., those which had been transcribed manually, with the aim of completing the results in the publication. This partition is quite different including new type of audiovisual contents, such as the, fiction genre, comedy and humor shows, live interviews during the recording session, among others. Consequently, new challenging acoustic conditions are included in the evaluation: movies with background music combined with synthetic speech, comedians disguising their voices, or different channels according to the microphones used in an interview without any post-production process, are some of the factors. Results showed that applying DAT did not reduce the WER.

As the evaluation of our systems was carried out under the conditions of the IberSpeech-RTVE 2020 Challenge (Lleida et al. (2020a)), we can now compare our work to other ASR systems participating in this challenge and thus also trained on this specific domain related to TV programs. As a first general comment, we can say that all the results for the developed ASR systems showed that end-to-end systems are still far from hybrid systems in the challenging conditions of RTVE 2020. Kocour et al. (2021) developed an end-to-end system based on the wav2letter architecture (Collobert et al. (2016)), which was not able to generalize very well on the acoustic conditions of the RTVE2020 database, reporting a WER of at least 13% higher than their best hybrid ASR system. Álvarez et al. (Alvarez et al. (2021)) presented a Quartznet-based (Kriman et al. (2020)) ASR implementation showing promising results due to the use of more than a hundred hours of speech data. Nevertheless, the results in terms of the WER from that system were 9% worse when compared to other hybrid systems they developed for the challenge. Furthermore, the acoustic conditions of databases provided in other international challenges, such as CHiME-6 (Watanabe et al. (2020)), have also been responsible for the poor performance of end-to-end ASR systems. Different approaches based on RNNs and transformers, along with the RNN-T and CTC-Attention (Andrusenko et al. (2020)) have been evaluated on one of the

CHiME partitions, concluding that speech-enhancement techniques contribute to the reduction of to the gap between end-to-end and hybrid systems. The difficulty of obtaining quality labeled speech leads to emergent machine-learning-based paradigms, such as Self-Supervised Learning (SSL), whose goal for ASR is to learn powerful speech representations from unlabeled examples (e.g., wav2vec 2.0 (Baevski et al. (2020))). Other recent works have tried to mitigate the effect of limited training data (Vyas et al. (2021)) or noisy environment conditions (Wang et al. (2021)). In the case of (Vyas et al. (2021)), the use of CTC and end-to-end LF-MMI to fine-tune a wav2vec 2.0 model showed similar performance even for out-of-domain and cross-lingual adaptation. Regarding (Wang et al. (2021)), the authors integrated SSL with contrastive learning from original–noisy speech pairs to model representations with noise robustness. Along the same lines, we proposed the application of DAT in our baseline end-to-end ASR systems as an alternative to obtain robust features invariant to the domain, i.e., different acoustic conditions of TV shows. It is a fact that DAT is beneficial for building robust embeddings.

Finally, regarding computational requirements for speech transcription, we considered that PyChain performs two main stages: a first decoding stage and a second four-gram rescoring stage. In terms of Real-Time factor, the results showed that time requirements depend on the characteristics of the test partitions. Less time is required to transcribe the best acoustic conditions and less challenging scenarios. This might make sense as far as the confusion of the graph model is less complex to transcribe accurately.

4.2 Conclusions

In this Section we present the main conclusions that can be drawn from the research results obtained after applying adversarial learning in the three speech applications under study.

To properly present our conclusions, we will discuss them while reviewing the specific objectives (O) which were outlined and defined in Section 1.2:

O1. Study of adversarial learning techniques.

The review of the state-of-the-art adversarial learning techniques leads us to conclude that the evolution of this technique as a type of adaptation towards a supervised method is a competitive approach to develop domain invariant features in speech applications. This conclusion is supported by an exhaustive search for DAT-related approaches. Table 2.1 includes a list of relevant references where we indicate the primary and secondary tasks, i.e., source and domain respectively, together with the improvement reported by each reference.

O2. Application of the adversarial learning techniques for the Assessment of

Obstructive Sleep Apnea from speech.

This objective has been clearly fulfilled. This is the first area of application that we worked on. Our principal findings contributed to the improvement of the OSA assessment in terms of accuracy. This is possible thanks to adversarial learning which helps to remove the effect of the sources of variability, such as the speaker's BMI, from the OSA-related features.

In a first experiment, we reported better results obtained when predicting AHI from i-vectors, MAE=13.45, $\rho = 0.32$, than from x-vectors, MAE=14.13, $\rho = 0.18$, can be explained as that the i-vectors were trained using a generative approach, trying to represent global speech variability, while x-vectors are discriminatively trained for a speaker identification task. Our study also showed the weaker impact of OSA on speech compared to the effect of speaker's age.

In a second experiment, we evaluate adversarial learning to classify between extreme OSA cases (AHI$\leq$ 10 and AHI$\geq$ 30) using BMI and AGE as adversarial domains. Positive results (an increase in accuracy from 69.39% to 76.60%) were only obtained when using DAT with BMI as the adversarial domain. These results seem to indicate that the impact of AGE in speech is more direct, while that of BMI is a subtler. In fact, medical research on anatomical pharyngeal and craniofacial abnormalities contrasting OSA patients to non-OSA snorers (therefore a somewhat symptomatic population as ours) has concluded that the shape of the pharyngeal lumen is more dependent on BMI than on the presence of OSA (Mayer et al. (1996)). This can explain the relevance of deriving speech features invariant to BMI.

O3. Application of the adversarial learning techniques for Speech Anonymization.

This third objective has also been accomplished. This is the second area of application where adversarial learning was integrated and evaluated. Our main contribution is the design and development of one of the best state-of-the-art anonymization methods according to the VoicePrivacy 2020 Challenge. Speech anonymization may not be a simple task depending on the final objective. The speaker identity is masked according to the objective results, taking into account the difficulty of maintaining speech intelligibility. We proposed adversarial learning to avoid revealing the speaker's identity and other characteristics (e.g., gender, accents) without varying the spoken content.

Our experimental results allow us to conclude that the proposed system, using an Autoencoder-Adversarial Network (AAN) to reconstruct the speaker x-vector, while removing private speaker characteristics, outperforms the Baseline of the VoicePrivacy 2020 Challenge in terms of privacy and utility for the majority of the evaluated attack scenarios. Compared with other submitted systems to the VoicePrivacy Challenge, our method achieved a high performance in terms of

privacy, i.e., EER is one of the highest values even improving the Baseline result, preserving the intelligibility, i.e., WER is practically the same as for the Baseline.

It is also important to mention the role of the trade-off parameter λ in the AAN model, as it can be used to control the balance between linguistic quality (i.e. reconstruction error) and the amount of speaker information removed.

Another interesting conclusion is that the best anonymization performance was achieved by the AAN with multiple adversarial domains, which explains the contribution of each speaker's characteristic adversarial domain to the suppression of speaker characteristic from the original x-vectors in the AAN.

O4. Application of adversarial learning techniques for Automatic Speech Recognition.

We also claim that this objective has been fulfilled. This corresponds to the last (but no less significant), application area that we worked on in this Thesis. Our contribution is related to the robustness of the ASR performance in case of adverse scenarios when recording TV programs.

Taking into account not only our experimental results, but also those provided by other systems participating in IberSpeech-RTVE 2020 Challenge, we can first conclude that end-to-end ASR systems still seems to be far from hybrid ASR models for complex tasks, such as TV transcription scenarios.

Looking for improvements in end-to-end ASR systems, we have proposed an adversarial learning approach integrated into an end-to-end LF-MMI model. In the proposed approach, the LF-MMI loss function is combined with an adversarial domain loss which aims making deep acoustic features invariant to the domain of the TV show characteristics. In that way we have been able to test the use of acoustic representations invariant to the domain of the TV show characteristics by using a Domain Adversarial Neural Network (DANN). Although, when including adversarial learning, we have only obtained a small improvement in WER, we believe that the combination of adversarial and self-supervised learning could be useful to achieve significant improvements in ASR systems through robust embeddings, even when trained without supervised data.

As an overall conclusion, we can say that this Thesis has shown how adversarial learning principles can be integrated into different deep learning architectures and for different speech applications. Although in some cases the improvements reported for adversarial learning are not very high, we hope that the different insights and approaches tested in every experimental scenario may be valuable for future research in this field.

4.3 Future works

The results achieved over the course of this Thesis, after applying adversarial learning to each field of application, led to a number of remarkable lines of research for the future:

- From the application of adversarial learning techniques for the assessment of Obstructive Sleep Apnea from speech, we should first point out that our OSA voices samples are not balanced. Most of the individuals are snorers reporting daytime sleepiness, there existing also a strong bias towards overweight and ages above 55 years old. Moreover, previous study (Solé-Casals et al. (2014)) reported the dependence on the recording position affecting the final performance of the model, due to the nature of the disorder. All these aspects related to the OSA voice samples could be considered for future research.

  Another interesting future line of research could be to study only lean or young subjects, because as shown in (Mayer et al. (1996)), it is only in this population where upper airway abnormalities explain a major part of the variance in AHI and are likely to play an important physiopathogenic role.

  Data augmentation techniques in OSA assessment could be addressed to overcome the main problem of the lack data for those less frequent cases. Other adversarial architectures such as Generative Adversarial Networks (GANs) to perform this issue, allowing us to extend our research to OSA female population, even considering the increase of female cases in our dataset, from which reliable results could be obtained.

- From the application of adversarial learning techniques for Speech Privacy, we could extend our current research to other speaker information, which was available on additional speech datasets, such as age or emotion. Thus, we could reduce the limitation of our work by suppressing only the information related to speaker identity, gender or accent.

  Another future line could be focused on the application of this framework to both encoded speech content and prosodic features, in a search for a better anonymization of the speech waveform by suppressing the speaker information while preserving the linguistic content. In addition, the exploration of generative models, such as variational autoencoders for the x-vector anonymization method can also contribute to speaker anonymization.

  One deep learning-based techniques to explore in the future is disentanglement. It is a powerful method to address the separation of information sources in speech. In this vein, it could be studied in the application fields we have worked on in the course of this Thesis.

Disentanglement has been used to separate subspaces from speech, such as speaking style, speaker's emotion, or speaker's gender. The application of this technique differs from domain adversarial training where the domain characteristics are subtracted from voice. Thus, the separation of multiple sources of variability allows retention of this information which can be valuable later.

- From the application of adversarial learning techniques for Automatic Speech Recognition in challenging scenarios, as a future work, we believe that exploring speech enhancement techniques could help to close the performance gap between hybrid and end-to-end ASR systems.

  Moreover, unsupervised machine learning methods (e.g., clustering) or automatic perceptual speech quality methods (e.g., PESQ) could contribute to a more accurate TV show classification prior to DAT.

  Finally, it is a fact that DAT is beneficial for building robust embeddings. In the same vein, we also believe that the combination of DAT with a self-supervised approach (wav2vec2.0) could be useful to achieve significant improvements in ASR systems through robust embeddings, even when trained without supervised data.

4.4 Framework

This PhD Thesis has been conducted in the GAPS Signal Processing Applications Group of the Signals, Systems and Radio-communications Department in the ETSIT of Universidad Politécnica de Madrid, under the supervision of Dr. Luis A. Hernández Gómez; and developed in Sigma Technologies SLU where Dr. Fernando Manuel Espinoza Cuadros has served as co-director of it.

All the contributions presented in this work have been funded by Universidad Politécnica de Madrid together with Sigma Technologies SLU, which offered their infrastructures and support to carry out all the needed tasks.

The first contribution (Section 3.1) was supported in part by the Spanish Ministry of Economy and Competitiveness and in part by the European Union (FEDER) as part of the TEC2015-68172-C2-2 (Deep & Subspace Speech Learning, DSSL) Project. This work was also possible through OSA database which was acquired for previous studies in collaboration with the Hospital Quirón Salud de Málaga.

The second contribution (Section 3.2) was supported by the UPM Research Grant RP2109550034

"Aplicación de Técnicas de Aprendizaje Automático sobre Señales y Datos no estructurados". This work was evaluated on the VoicePrivacy 2020 Challenge, a satellite workshop in Interspeech, where the organizers provided the participants with all the data and the evaluation plan.

The third contribution (Section 3.3) does not receive any external funding. In terms of the data, RTVE2020 database has been seconded by RTVE Corporation and Universidad de Zaragoza aiming to contribute to the development of the Speech Technologies in Spanish. This work was also evaluated on the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge, one of the most important ASR evaluations for Iberian languages. In addition, it should be noted the hard labor accomplished by Sigma Technologies supervising the data to make reliable transcriptions to train ASR models accurately.

All the works along this PhD Thesis were reached from June 2018 to January 2022. During these years, the PhD candidate has worked toward this PhD degree as part-time job while has been with the Head of Speech Technologies at Sigma Technologies. As a consequence, he has taken part in several activities, including R&D projects, applications and solutions based on Artificial Intelligence, which have given a valuable experience in several aspects of this Thesis.

Bibliography

Adi, Y., Zeghidour, N., Collobert, R., Usunier, N., Liptchinsky, V. and Synnaeve, G. To Reverse the Gradient or Not: An Empirical Comparison of Adversarial and Multi-task Learning in Speech Recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, pages 3742–3746. Institute of Electrical and Electronics Engineers Inc., 2019. ISBN 9781479981311. ISSN 15206149. 3, 29

Alvarez, A., Arzelus, H., Torre, I. G. and González-Docasal, A. The vicomtech speech transcription systems for the albayzín-rtve 2020 speech to text transcription challenge. In *IberSPEECH*. 2021. 85

Andrusenko, A., Laptev, A. and Medennikov, I. Towards a competitive end-to-end speech recognition for chime-6 dinner party transcription. *arXiv preprint arXiv:2004.10799*, 2020. 85

Baevski, A., Zhou, H., Mohamed, A. and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020. 86

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. and Bengio, Y. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE, 2016. 34

Bahmaninezhad, F., Zhang, C. and Hansen, J. H. Convolutional neural network based speaker de-identification. In *Odyssey*, pages 255–260. 2018. 26

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, vol. 79(1), pages 151–175, 2010. 2

Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, vol. 19, 2006. 2

Botelho, M. C., Trancoso, I., Abad, A. and Paiva, T. Speech as a biomarker for obstructive sleep apnea detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5851–5855. IEEE, 2019. 18

Brasser, F., Frassetto, T., Riedhammer, K., Sadeghi, A.-R., Schneider, T. and Weinert, C. Voiceguard: Secure and private speech processing. In *Interspeech*, vol. 18, pages 1303–1307. 2018. 25

Caruana, R. Multitask learning. *Machine learning*, vol. 28(1), pages 41–75, 1997. 2

Chan, W., Jaitly, N., Le, Q. and Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE, 2016. 34

Chung, J. S., Nagrani, A. and Zisserman, A. VoxCeleb2: Deep Speaker Recognition. In *Interspeech 2018*, pages 1086–1090. ISCA, ISCA, 2018. 31

Chung, Y.-A., Hsu, W.-N., Tang, H. and Glass, J. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240*, 2019. 2

Cohen-Hadria, A., Cartwright, M., McFee, B. and Bello, J. P. Voice anonymization in urban sound recordings. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2019. 25

Collobert, R., Puhrsch, C. and Synnaeve, G. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016. 85

Denisov, P., Vu, N. T. and Font, M. F. Unsupervised Domain Adaptation by Adversarial Learning for Robust Speech Recognition. Technical report, Institute for Natural Language Processing, University of Stuttgart, Germany, 2018. 16, 35

Espinoza-Cuadros, F., Fernández-Pozo, R., Toledano, D. T., Alcázar-Ramírez, J. D., López-Gonzalo, E. and Hernández-Gómez, L. A. Speech Signal and Facial Image Processing for Obstructive Sleep Apnea Assessment. *Computational and Mathematical Methods in Medicine*, vol. 2015, pages 1–13, 2015. ISSN 1748-670X. 19

Espinoza-Cuadros, F., Fernández-Pozo, R., Toledano, D. T., Alcázar-Ramírez, J. D., López-Gonzalo, E. and Hernández-Gómez, L. A. Reviewing the connection between speech and obstructive sleep apnea. *Biomedical engineering online*, vol. 15(1), page 20, 2016. ISSN 1475-925X. 19, 22, 23, 80

Espinoza Cuadros, F. M. *Study of speech and craniofacial features in obstructive sleep apnea patients*. PhD thesis, Telecomunicacion, 2018. 19

Espinoza-Cuadros, F. M., Perero-Codosero, J. M., Antón-Martín, J. and Hernández-Gómez, L. A. Speaker de-identification system using autoencoders and adversarial training. 2020. 6

Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N. W. D. and Bonastre, J.-F. Speaker anonymization using x-vector and neural waveform models. *ArXiv*, vol. abs/1905.13561, 2019. 26, 27, 28

Fernández Pozo, R., Blanco Murillo, J. L., Hernández Gómez, L., López Gonzalo, E., Alcázar Ramírez, J. and Toledano, D. T. Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pages 1–11, 2009. 18, 19

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 14

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., Dogan, U., Kloft, M., Orabona, F., Tommasi, T. and Ganin, A. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, vol. 17, pages 1–35, 2016. 2, 13, 16, 29

Gontier, F., Lagrange, M., Lavandier, C. and Petiot, J.-F. Privacy aware acoustic scene synthesis using deep spectral feature inversion. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 886–890. IEEE, 2020. 25

González-Rodríguez, J., Gil, J., Pérez, R. and Franco-Pedroso, J. What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials. In *Proceedings of Odyssey*, pages 33–40. 2014. 19

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, vol. 27, 2014. 12

Graves, A. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012. 34

Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. 2006. 34

Hadian, H., Sameti, H., Povey, D. and Khudanpur, S. End-to-end speech recognition using lattice-free mmi. In *Interspeech*, pages 12–16. 2018. 34

Hashimoto, K., Yamagishi, J. and Echizen, I. Privacy-preserving sound to degrade automatic speaker verification performance. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5500–5504. IEEE, 2016. 25

ISO/IEC JTC1 SC27 Security Techniques. ISO/IEC 24745:2011. Information Technology-Security Techniques-Biometric Information Protection. Technical report, International Organization for Standardization, 2011. 24

Janbakhshi, P. and Kodrasi, I. a. Supervised speech representation learning for parkinson's disease classification. In *Speech Communication; 14th ITG Conference*, pages 1–5. VDE, 2021. 18

Jin, Q., Toth, A. R., Schultz, T. and Black, A. W. Speaker de-identification via voice transformation. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 529–533. IEEE, 2009. 26

Justin, T., Štruc, V., Dobrišek, S., Vesnicer, B., Ipšić, I. and Mihelič, F. Speaker de-identification using diphone recognition and speech synthesis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 04, pages 1–7. 2015. 26

Kocour, M., Cámbara, G., Luque, J., Bonet, D., Farrús, M., Karafiát, M., Veselỳ, K. and Ĉernockỳ, J. Bcn2brno: Asr system fusion for albayzin 2020 speech to text challenge. *arXiv preprint arXiv:2101.12729*, 2021. 85

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J. and Zhang, Y. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE, 2020. 85

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86(11), pages 2278–2324, 1998. 14, 16

Leroy, D., Coucke, A., Lavril, T., Gisselbrecht, T. and Dureau, J. Federated learning for keyword spotting. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6341–6345. IEEE, 2019. 25

Li, H., Tu, M., Huang, J., Narayanan, S. and Georgiou, P. Speaker-invariant affective representation learning via adversarial training. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7144–7148. IEEE, 2020. 18

Li, Y., Swersky, K. and Zemel, R. Learning unbiased features. *arXiv preprint arXiv:1412.5244*, 2014. 16

Lleida, E., Ortega, A., Miguel, A., Bazán-Gil, V., Pérez, C., Gómez, M. and de Prada, A. Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media. *Applied Sciences*, vol. 9(24), page 5412, 2019. 7

Lleida, E., Ortega, A., Miguel, A., Bazán-Gil, V., Pérez, C., Gómez, M. and de Prada, A. Albayzin evaluation: Iberspeech-rtve 2020 speech to text transcription challenge. http://catedrartve.unizar.es/reto2020/EvalPlan-S2T-2020-v1.pdf, 2020a. [Online]. 35, 37, 38, 85

Lleida, E., Ortega, A., Miguel, A., Bazán-Gil, V., Pérez, C., Gómez, M. and de Prada, A. Rtve2020 database description. http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf, 2020b. [Online]. 37

López-González, E., Caminero, J., Cortázar, I. and Hernández-Gómez, L. A. Improvement on connected numbers recognition using prosodic information. In *ICSLP*. 1998. 22

Lorenzo-Trueba, J., Fang, F., Wang, X., Echizen, I., Yamagishi, J. and Kinnunen, T. H. Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data. *ArXiv*, vol. abs/1803.00860, 2018. 26

Louizos, C., Swersky, K., Li, Y., Welling, M. and Zemel, R. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015. 16

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, vol. 9(Nov), pages 2579–2605, 2008. 80, 81

Magariños, C., Lopez-Otero, P., Docio-Fernandez, L., Rodriguez-Banga, E., Erro, D. and Garcia-Mateo, C. Reversible speaker de-identification using pre-trained transformation functions. *Computer Speech Language*, vol. 46, pages 36–52, 2017. ISSN 0885-2308. 26

Mayer, P., Pepin, J. L., Bettega, G., Veale, D., Ferretti, G., Deschaux, C. and Lévy, P. Relationship between body mass index, age and upper airway measurements in snorers and sleep apnoea patients. *European Respiratory Journal*, vol. 9(9), pages 1801–1809, 1996. 87, 89

Mcadams, S. Spectral fusion, spectral parsing and the formation of auditory images. 1984. 29

Mdhaffar, S., Bonastre, J.-F., Tommasi, M., Tomashenko, N. and Estève, Y. Retrieving speaker information from personalized acoustic models for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6767–6771. IEEE, 2022. 25

Meng, Z., Li, J., Chen, Z., Zhao, Y., Mazalov, V., Gang, Y. and Juang, B. H. Speaker-Invariant Training Via Adversarial Learning. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pages 5969–5973. Institute of Electrical and Electronics Engineers Inc., 2018. ISBN 9781538646588. ISSN 15206149. 3, 17, 29

Montero-Benavides, A., Blanco-Murillo, J. L., Fernández, A., Fernández-Pozo, R., Toledano, D. T. and Hernández-Gómez, L. A. Using HMM to detect speakers with severe obstructive sleep apnoea syndrome. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 121–128. Springer, 2012. 19

Moreno-Bilbao, M. A., Poig, D., Bonafonte-Cávez, A., Lleida, E., Llisterri, J., Mariño-Acebal, J. B. and Nadeu-Camprubí, C. Albayzin speech database: Design of the phonetic corpus. In *EUROSPEECH, 1993.*, pages 175–178. . EUROSPEECH, 1993. 22

Nagrani, A., Chung, J. S. and Zisserman, A. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Interspeech 2017*, pages 2616–2620. ISCA, ISCA, 2017. 31

Panayotov, V., Chen, G., Povey, D. and Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2015-August, pages 5206–5210. Institute of Electrical and Electronics Engineers Inc., 2015. ISBN 9781467369978. ISSN 15206149. 31

Pascual, S., Ravanelli, M., Serra, J., Bonafonte, A. and Bengio, Y. Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv preprint arXiv:1904.03416*, 2019. 2

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037. 2019. 34

Pathak, M. A., Raj, B., Rane, S. D. and Smaragdis, P. Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise. *IEEE signal processing magazine*, vol. 30(2), pages 62–74, 2013. 25

Patino, J., Tomashenko, N., Todisco, M., Nautsch, A. and Evans, N. Speaker anonymisation using the mcadams coefficient. *arXiv preprint arXiv:2011.01130*, 2020. 26

Peddinti, V., Povey, D. and Khudanpur, S. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 3214–3218. 2015. 35

Perero-Codosero, J. M., Espinoza-Cuadros, F., Anton-Martin, J., Barbero-Alvarez, M. A. and Hernandez, L. A. Modeling Obstructive Sleep Apnea voices using Deep Neural Network Embeddings and Domain-Adversarial Training. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–1, 2019. ISSN 1932-4553. 20

Perero-Codosero, J. M., Espinoza-Cuadros, F. M. and Gómez, L. A. H. Sigma-upm asr systems for the iberspeech-rtve 2020 speech-to-text transcription challenge. In *IberSPEECH*. 2021. 6

Perero-Codosero, J. M., Espinoza-Cuadros, F. M. and Hernández-Gómez, L. A. X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. *Computer Speech & Language*, vol. 74, page 101351, 2022. 29

Pobar, M. and Ipšić, I. Online speaker de-identification using voice transformation. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1264–1267. 2014. 26

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y. and Khudanpur, S. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755. 2016. 34

Qian, J., Du, H., Hou, J., Chen, L., Jung, T., Li, X.-Y., Wang, Y. and Deng, Y. Voicemask: Anonymize and sanitize voice input on mobile devices. *arXiv preprint arXiv:1711.11460*, 2017. 25

Rohdin, J., Stafylakis, T., Silnova, A., Zeinali, H., Burget, L. and Plchot, O. Speaker verification using end-to-end adversarial language adaptation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6006–6010. IEEE, 2019. 18

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B. and Hall, P. English Conversational Telephone Speech Recognition by Humans and Machines. In *Interspeech 2017*, vol. 2017-August, pages 132–136. ISCA, ISCA, 2017. 29

Serdyuk, D., Audhkhasi, K., Brakel, P., Ramabhadran, B., Thomas, S. and Bengio, Y. Invariant representations for noisy speech recognition. *arXiv preprint arXiv:1612.01928*, 2016. 3, 15, 16, 29, 35

Shao, Y., Wang, Y., Povey, D. and Khudanpur, S. Pychain: A fully parallelized pytorch implementation of lf-mmi for end-to-end asr. *arXiv preprint arXiv:2005.09824*, 2020. 34, 85

Shinohara, Y. Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, pages 2369–2372. International Speech and Communication Association, 2016. 2, 14, 16, 29, 35

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018. 26, 28

Solé-Casals, J., Munteanu, C., Martín, O. C., Barbé, F., Queipo, C., Amilibia, J. and Durán-Cantolla, J. Detection of severe obstructive sleep apnea through voice analysis. *Applied Soft Computing*, vol. 23, pages 346–354, 2014. ISSN 15684946. 18, 19, 80, 89

Srivastava, B. M. L., Bellet, A., Tommasi, M. and Vincent, E. Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion? *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, pages 3700–3704, 2019. 29

Sun, S., Yeh, C.-F., Hwang, M.-Y., Ostendorf, M. and Xie, L. Domain Adversarial Training for Accented Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4854–4858. IEEE, 2018. ISBN 978-1-5386-4658-8. 3, 14, 17, 29, 35

Toledano, D., Hernandez Gomez, L. and Grande, L. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, vol. 11(6), pages 617–625, 2003. ISSN 1063-6676. 22, 37

Tomashenko, N., Mdhaffar, S., Tommasi, M., Estève, Y. and Bonastre, J.-F. Privacy attacks for automatic speech recognition acoustic models in a federated learning framework. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6972–6976. IEEE, 2022. 25

Tomashenko, N., Mohan, B., Srivastava, L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G. and Todisco, M. The VoicePrivacy 2020 Challenge Evaluation Plan. pages 1–17, 2020a. 33

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G. and Todisco, M. Introducing the voiceprivacy initiative. 2020b. 8, 25, 27, 31, 32, 80

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G., Todisco, M., Maouche, M., O'brien, B. and Chanclu, A. Challenge setup and results the voiceprivacy 2020 challenge odyssey 2020. 2020c. 8

Tsuchiya, T., Tawara, N., Ogawa, T. and Kobayashi, T. Speaker Invariant Feature Extraction for Zero-Resource Languages with Adversarial Learning. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pages 2381–2385. Institute of Electrical and Electronics Engineers Inc., 2018. ISBN 9781538646588. ISSN 15206149. 3, 17, 29

Vyas, A., Madikeri, S. and Bourlard, H. Comparing ctc and lfmmi for out-of-domain adaptation of wav2vec 2.0 acoustic model. *arXiv preprint arXiv:2104.02558*, 2021. 86

Wang, X., Takaki, S. and Yamagishi, J. Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pages 402–415, 2020. 27

Wang, Y., Li, J., Wang, H., Qian, Y., Wang, C. and Wu, Y. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. *arXiv preprint arXiv:2110.04934*, 2021. 86

Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D. et al. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*, 2020. 85

Yamagishi, J., Veaux, C., MacDonald, K. et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019. 31

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z. and Wu, Y. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Interspeech 2019*, pages 1526–1530. ISCA, ISCA, 2019. 31

Zhang, S.-X., Gong, Y. and Yu, D. Encrypted speech recognition using deep polynomial networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5691–5695. IEEE, 2019. 25

Zigel, Y., Tarasiuk, A. and Goldshtein, E. Analysis of speech signals among obstructive sleep apnea patients. In *Electrical and Electronics Engineers in Israel, 2008. IEEEI 2008. IEEE 25th Convention of*, pages 760–764. IEEE, 2008. 18, 19