Investigating Adaptation and Transfer Learning for End-to-End Spoken Language Understanding from Speech

Natalia Tomashenko, Antoine Caubrière, Yannick Estève

INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Investigating Adaptation and Transfer Learning for End-to-End Spoken Language Understanding from Speech

Natalia Tomashenko[1], *Antoine Caubrière*[2], *Yannick Estève*[1]

[1]LIA, University of Avignon, France
[2]LIUM, University of Le Mans, France

natalia.tomashenko@univ-avignon.fr, antoine.caubriere@univ-lemans.fr, yannick.esteve@univ-avignon.fr

Abstract

This work investigates speaker adaptation and transfer learning for spoken language understanding (SLU). We focus on the direct extraction of semantic tags from the audio signal using an end-to-end neural network approach. We demonstrate that the learning performance of the target predictive function for the semantic slot filling task can be substantially improved by speaker adaptation and by various knowledge transfer approaches. First, we explore speaker adaptive training (SAT) for end-to-end SLU models and propose to use zero pseudo i-vectors for more efficient model initialization and pretraining in SAT. Second, in order to improve the learning convergence for the target semantic slot filling (SF) task, models trained for different tasks, such as automatic speech recognition and named entity extraction are used to initialize neural end-to-end models trained for the target task. In addition, we explore the impact of the knowledge transfer for SLU from a speech recognition task trained in a different language. These approaches allow to develop end-to-end SLU systems in low-resource data scenarios when there is no enough in-domain semantically labeled data, but other resources, such as word transcriptions for the same or another language or named entity annotation, are available.

Index Terms: adaptation, end-to-end models, named entity recognition, automatic speech recognition, spoken language understanding, deep neural networks, semantic slot filling

1. Introduction

Traditional SLU systems consist of several components: (1) an automatic speech recognition (ASR) system that transcribes acoustic speech signal into word sequences and (2) a natural language understanding (NLU) system which predicts, given the output of the ASR system, named entities, semantic or domain tags, and other language characteristics depending on the considered task. In classical approaches, these two systems are usually built and optimized independently.

In the recent years, there has be a great interest of the research community in end-to-end systems for various speech and language technologies, such as ASR [1, 2, 3, 4], text-to-speech synthesis [5], machine translation [6], speaker verification [7] and many others. A few recent papers [8, 9, 10, 11, 12] present ASR-free end-to-end approaches for SLU tasks and show promising results. These methods aim to learn SLU models from acoustic signal without intermediate text representation. Paper [12] proposed an audio-to-intent architecture for semantic classification in dialog systems. An encoder-decoder framework [13] is used in paper [10] for domain and intent classification, and in [9] for domain, intent, and argument recognition. A different approach based on the model trained with the connectionist temporal classification (CTC) criterion [14] was proposed in [11] for named entity recognition (NER) and slot filling, and it is the closest to the current work.

These methods are motivated by the following factors: (1) possibility of better information transfer from the speech signal due to the joint optimization on the final objective function, and, in particular, leveraging errors from the ASR system and focusing on the most important information; and (2) simplification of the overall system; getting rid of some components, such as pronunciation lexicon, etc.

In this paper, we focus on the two SLU tasks: named entity recognition (NER) and semantic slot filling (SF). The target task in this paper is SF, and we use the NER task as an auxiliary task for transfer learning. The aim of this work is to explore the efficiency of speaker adaptation and knowledge transfer for end-to-end SLU models.

The rest of the paper is organized as follows. Section 2 presents a review on speaker adaptation for end-to-end models and the proposed adaptation approach. Section 3 introduces the transfer learning approaches that we investigate in the current work. Sections 4 describes the experimental setup and results. Finally, the conclusions are given in Section 5.

2. Speaker adaptation

Differences between training and testing conditions may significantly reduce recognition accuracy in ASR systems and degrade performance of other speech-related technologies. Adaptation is an efficient way to reduce the mismatches between the models and the data from a particular speaker or channel. For many decades, acoustic model adaptation has been an essential component of any state-of-the-art ASR system. For end-to-end approaches, speaker adaptation is less studied, and most of the first end-to-end ASR systems do not use any speaker adaptation and are built on spectrograms [1, 3] or filterbank features [4, 15]. However, some recent works [16, 17, 18, 19] demonstrated the effectiveness of speaker adaptation for end-to-end models. Various feature-space speaker adaptation techniques, such as i-vectors [20, 21], feature-space maximum linear regression (fMLLR) [22] and maximum a posteriori (MAP) adaptation [23] using GMM-derived features [24] were investigated in [16] for bidirectional long short term memory (BLSTM) recurrent neural network based acoustic models (AMs) trained with the CTC objective function. In [17], an auxiliary feature based adaptation in the form of a sequence summary network is studied for end-to-end encoder-decoder models. Adaptation for multi-channel end-to-end encoder-decoder ASR model was explored in [18]. Kullback-Leibler divergence (KLD) regularization and multi-task learning (MTL) was investigated in [19] for CTC models.

For SLU tasks, there is also an emerging interest in the

http://dx.doi.org/10.21437/Interspeech.2019-2158

end-to-end models which have a speech signal as input. Thus, acoustic, and particularly speaker, adaptation for such models can play an important role in improving the overall performance of these systems. However, to our knowledge, there is no research on speaker adaptation for end-to-end SLU models, and the existing works do not use any speaker adaptation. In [8], Mel frequency cepstral coefficient (MFCC) features were used for an ASR-free end-to-end NLU model for dialog systems. Papers [9] and [10] use log-Mel filterbanks in encoder-decoder based end-to-end approaches: [9] – for domain, intent, and argument prediction; and [10] – for intent and domain classification. In [11], end-to-end CTC-based systems for NER and SF were built on spectograms. For semantic classification, an ASR-free system was built in [12] on log-spectrum features.

One of the main objectives of this work is to explore speaker adaptation for end-to-end SLU. For experiments in this paper, we apply i-vector based speaker adaptation [21, 20]. I-vectors can capture the relevant information about the speaker in a low-dimensional fixed-length representation [21].

2.1. Integration of i-vectors into end-to-end models

The proposed way of integration of i-vectors into the end-to-end model architecture is shown in Figure 1. Speaker i-vectors are appended to the outputs of the last (second) convolutional layer, just before the first recurrent (BLSTM) layer. In our preliminary experiments (not reported in this paper), we also tried other ways of i-vector integration (in particularly, to append to upper or to lower layers, or to several layers) and found out the chosen configuration is the most efficient. We do not append i-vectors to the input layer, because the first two layers in our model are convolutional, incorporation of auxiliary features is not straightforward since i-vectors do not have the same time and frequency locality properties as input acoustic features. Thus, incorporation of auxiliary features to a convolutional layer makes a system more complex [25].

In this paper, we experiment with two ways of speaker adaptive training. For better initialization, we first propose to train a model with *zero pseudo i-vectors* (all values are equal to 0). Then, we use this pretrained model and fine-tune a new model on the same data but with the real i-vectors. This approach was inspired by [26], where an idea of using zero auxiliary features during pretraining was implemented for language models. For comparison purpose, we also train a model directly on real i-vectors without pretraining with zero i-vectors.

3. Transfer learning for end-to-end SLU

Transfer learning is a popular and efficient method to improve the learning performance of the target predictive function using knowledge from a different source domain [27]. It allows to train a model for a given target task using available out-of-domain source data, and hence to avoid an expensive data labeling process, which is especially useful in case of low-resource scenarios.

In this paper, the target task is semantic slot filling (SF). We investigate the effectiveness of the transfer learning paradigm for various source domains and tasks:

1. ASR
   (a) in the target language;
   (b) in the out-of-domain language;
2. NER in the target language;
3. Slot filling (SF).

Similarly to point 1(a), transfer learning from ASR to SF in the target language was used in [11], however the gain of this approach was not reported.

For all the tasks, we used similar model architectures (Section 4.2 and Figure 1). The difference is in the text data preparation and output targets. For training ASR systems, the output targets correspond to alphabetic characters and a 'blank' (no label) symbol. For NER tasks, the output targets include all the ASR targets and targets corresponding to named entity tags. We have several symbols corresponding to named entities (in the text these characters are situated before the beginning of a named entity, which can be a single word or a sequence of several words) and a one tag corresponding to the end of the named entity, which is the same for all named entities. Similarly, for SF tags, we use targets corresponding to the semantic concept tags and one tag corresponding to the end of the given concept.

Transfer learning is realized through the chain of consequence model training on different tasks. For example, we can start from training an ASR model on audio data and corresponding text transcriptions. Then, we change the softmax layer in this model by replacing the targets with the SF targets and continue training on the corpus annotated with semantic tags. Further in the paper, we denote this type of chain as $ASR \rightarrow SF$. Models in this chain can be trained on different corpora, that can make this method especially useful in low-resource scenarios when we do not have enough semantically annotated data to train an end-to-end model, but have sufficient amount of data annotated with more general concepts or only transcribed data. Details on the use of this approach are presented in [28].

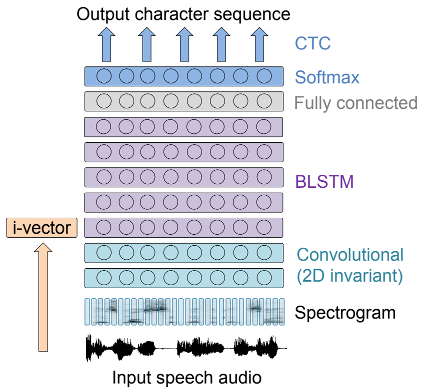

Figure 1: *Universal end-to-end deep neural network model architecture for ASR, NER and SF tasks. Depending on the current task, the set of the output characters (targets) consists of: (1) 43 characters for French ASR and 28 – for English ASR; (2) 43+9=52 – for NER; and (3) 43+87=130 – for SF.*

Table 1: *Corpus statistics for ASR, NER and SF tasks.*

Task	Corpora	Size, h	# Speakers
ASR train	EPAC [29], ESTER 1,2 [30], ETAPE [31], REPERE [32], DECODA [33], MEDIA [34] PORTMEDIA [35]	404.6	12518
NER train	EPAC [29], ESTER 1,2 [30], ETAPE [31], REPERE [32]	323.8	7327
SF train	MEDIA [34] (train),	16.1	727
	PORTMEDIA [35] (train)	7.2	257
SF dev	MEDIA [34] (dev)	1.7	79
SF test	MEDIA [34] (test)	4.8	208

Table 2: *Results on the MEDIA test dataset for* ***speaker independent*** *end-to-end SF models trained with different transfer learning approaches. Results are given in terms of F-measure (F), CER and CVER metrics (%); ΔCVER denotes relative error reduction for CVER in comparison with the baseline model (#1). CER and CVER are reported with 95% confidence intervals shown in gray.* ***T*** *corresponds to the* ***T****arget task when a model is trained on MEDIA train data and* ***A*** *denotes the* ***A****uxiliary task when the model is trained on MEDIA+PORTMEDIA training data;* ***F*** *and* ***E*** *refer to the languages:* ***F****rench and* ***E****nglish; and "*" means a starred mode.*

		Without LM				With LM			
#	**Training chain**	**F**	**CER**	**CVER**	**ΔCVER**	**F**	**CER**	**CVER**	**ΔCVER**
1	SF_T	72.5	39.4 ± 1.0	52.7 ± 1.0	baseline	77.6	34.0 ± 1.0	39.7 ± 1.0	baseline
2	SF_A	73.2	39.0 ± 1.0	50.1 ± 1.1	4.9	77.9	33.8 ± 1.0	38.3 ± 1.0	3.5
3	$SF_A \to SF_T$	77.4	33.9 ± 1.0	44.9 ± 1.0	14.8	81.2	29.4 ± 1.0	34.3 ± 1.0	13.6
4	$ASR_E \to SF_A \to SF_T$	81.3	28.4 ± 0.9	37.3 ± 1.0	29.2	84.0	25.2 ± 0.9	29.7 ± 1.0	25.2
5	$ASR_F \to SF_A \to SF_T$	85.9	21.7 ± 0.9	28.4 ± 0.9	46.1	88.3	18.7 ± 0.8	22.8 ± 0.9	42.6
6	$NER \to SF_A \to SF_T$	86.4	20.9 ± 0.9	27.5 ± 0.9	47.8	88.0	18.9 ± 0.8	23.1 ± 0.9	41.8
7	$ASR_F \to SF_A \to SF_T^*$	85.9	21.2 ± 0.9	27.9 ± 0.9	47.1	88.6	**17.2** ± 0.8	**21.6** ± 0.9	45.6
	state-of-the-art [36]						19.9	25.1	

Table 3: ***Speaker adaptation*** *results on the MEDIA test dataset for end-to-end SF models trained with different transfer learning approaches (following the same numeration as in Table 2).* ***iv***$_0$ *corresponds to zero pseudo i-vector pretraining as described in Section 2.1, and* ***iv*** *is a standard using of i-vectors (without pretraining); ΔCER, ΔCVER (%) denote relative error reduction of CER and CVER for: (1) SAT models with the proposed zero pseudo i-vector pretraing with respect to the speaker independent (SI) models (see Table 2) "SI vs SAT with* ***iv***$_0$*"; and (2) SAT models with the proposed zero pseudo i-vector pretraing with respect to the SAT models with convention training using i-vectors: "SAT with* ***iv*** *vs* ***iv***$_0$*".*

	Without LM				With LM							
	iv		**iv$_0$**		**iv**		**iv$_0$**		(1) **SI vs SAT with iv$_0$**		(2) **SAT with iv vs iv$_0$**	
#	**CER**	**CVER**	**CER**	**CVER**	**CER**	**CVER**	**CER**	**CVER**	**ΔCER**	**ΔCVER**	**ΔCER**	**ΔCVER**
1	38.0	50.9	32.2	43.1	32.5	37.4	28.1	33.0	17.4	16.9	13.5	11.8
2	40.8	50.9	30.3	40.2	34.0	38.2	26.8	31.4	20.7	18.0	21.2	17.8
3	32.2	42.6	28.1	37.2	28.4	33.2	24.5	29.4	16.7	14.3	13.7	11.5
4	25.7	34.5	24.6	32.6	23.0	27.5	22.0	26.8	12.7	9.8	4.4	2.6
5	20.2	26.6	19.4	25.4	18.2	22.5	17.8	21.9	4.8	4.0	2.2	2.7
6	20.6	27.4	19.5	26.0	19.0	22.9	18.0	22.0	4.8	4.8	5.3	3.9
7	19.3	26.8	**18.8**	**25.5**	16.6	21.1	**16.4**	**20.8**	4.7	3.7	1.2	1.4

4. Experiments

4.1. Data

Several publicly available corpora have been used for experiments (see Table 1).

4.1.1. ASR data

The corpus for ASR training was composed of corpora from various evaluation campaigns in the field of automatic speech processing for French, as shown in Table 1. The EPAC [29], ESTER 1,2 [30], ETAPE [31], REPERE [32] contain transcribed speech in French from TV and radio broadcasts. These data were originally in the microphone channel and for experiments in this paper were downsampled from 16kHz to 8kHz, since the test set for our main target task (SF) consists of telephone conversations. The DECODA [33] corpus is composed of dialogues from the call-center of the Paris transport authority. The MEDIA [34, 37] and PORTMEDIA [35] are corpora of dialogues simulating a vocal tourist information server. The target language in all experiments is French. For experiments with transfer learning from ASR built in a different source language (English in our case) to SF in the target language, we used the TED-LIUM corpus [38]. This publicly available dataset contains 1495 TED talks in English that amount to 207 hours speech data from 1242 speakers, 16kHz. For experiments, we downsampled the audio data to 8kHz.

4.1.2. NER data

To train the NER system, we used the following corpora: EPAC, ESTER 1,2, ETAPE, and REPERE. These corpora contain speech with text transcriptions and named entity annotation. The named entity annotation is performed following the methodology of the Quaero project [39]. The taxonomy is composed of 8 main types: *person, function, organization, location, product, amount, time*, and *event*. Each named entity can be a single word or a sequence of several words. The total amount of annotated data is 112 hours. Based on this data, a classical NER system was trained using *NeuroNLP2*[1] to automatically extract named entities for the rest 212 hours of the training corpus. This was done in order to increase the amount of the training data for NER (as proposed in [11]). Thus, the total amount of audio data to train the NER system is about 324 (112+212) hours.

4.1.3. SF data

The following two French corpora, dedicated to semantic extraction from speech in a context of human/machine dialogues, were used in the current experiments: MEDIA and PORTMEDIA (see Table 1). The corpora have manual transcription and conceptual annotation. A concept is defined by a label and a value, for example with the concept *date*, the value *2001/02/03* can be associated [40, 34]. The MEDIA corpus is related to the hotel booking domain, and its annotation contains 76 semantic tags: *room number, hotel name, location, date, room equipment*, etc. The PORTMEDIA corpus is related to the theater ticket reservation domain and its annotation contains 35 semantic tags which are very similar to the tags used in the MEDIA corpus. For joint training on these corpora, we used a combined set of 86 semantic tags.

[1]https://github.com/XuezheMax/NeuroNLP2

4.2. Models

The neural architecture is similar to the Deep Speech 2 [3] for ASR. The two major differences in comparison with the original architecture are the following. First, we integrated speaker adaptation into this system based on i-vectors as shown in Figure 1 and proposed in Section 2.1. Second, in this paper, the tasks include NER and SF, therefore when we train neural networks for these tasks, the output sequence besides the alphabetic characters also contains special characters corresponding to named entities or semantic tags. A spectrogram of power normalized audio clips calculated on 20ms windows is used as the input features for the system. As shown in Figure 1, it is followed by two 2D-invariant (in the time and-frequency domain) convolutional layers, and then by five BLSTM layers with sequence-wise batch normalization [41]. A fully connected layer is applied after BLSTM layers, and the output layer of the neural network is a softmax layer. The model is trained using the CTC loss function [14]. We used the *deepspeech.torch* implementation[2] for training speaker independent models, and our modification of this implementation to integrate speaker adaptation. The open-source *Kaldi* toolkit [42] was used to extract 100-dimensional speaker i-vectors.

4.3. Results

Performance was evaluated in terms of *F-measure*, *concept error rate* (CER) and *concept value error rate* (CVER). A 4-gram LM with an about 4K word vocabulary built on French text data of the training corpus (including the semantic tags from MEDIA and PORTMEDIA training corpora) was used for evaluation.

Results for different training chains for speaker-independent (SI) models are given in Table 2. The first line SF_T shows the baseline result on the test MEDIA dataset for the SF task, when a model was trained directly on the target task using in-domain data for this task (the training part of the MEDIA corpus). The second line SF_A corresponds to the case when the model was trained on the *auxiliary* SF task, where targets were the same, but the training corpus was comprised of two corpora: the target corpus MEDIA and an additional corpus PORTMEDIA. The rest lines in the table correspond to different training chains described in Section 3. In $\#4$, we can see a chain that starts from training an ASR model for English. We can observe that using a pretrained ASR model from a different language can significantly (16.9% of relative CVER reduction, in case when no LM is used) improve the performance of the SF model (#4 vs #3). Using an ASR model trained in French (#5) provides better improvement: 36.7% of relative CVER reduction (#5 vs #3). When we start the training process from a NER model (#6) we can observe similar results. In #7, symbol "*" corresponds to a *starred mode* [11] where during the training, a new symbol "*" was added for targets, while in the texts all irrelevant words (according to the current task) were replaced by this character in order to make the learning process more focused on the target words and tags and to ignore less relevant information. This means that word sequence occurrences that do not appear within a concept are replaced by a star. For this mode, we also used a corresponding 4-gram LM which was built on the texts including "*". This SF model provides the best result when the LM is used for decoding. In terms of CER and CVER metrics, three last models (#5, #6 and #7), outperform the best published result (shown in the last line of the table) for SF for the MEDIA test task when the concept extraction was based on the ASR output.

[2] https://github.com/SeanNaren/deepspeech.pytorch

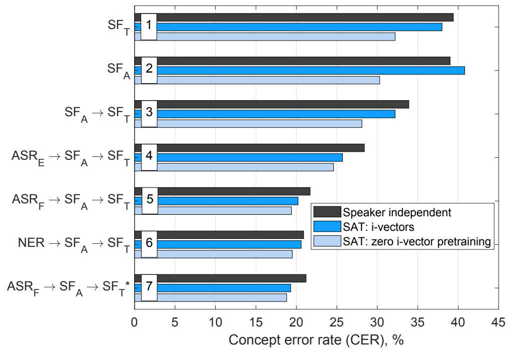

Figure 2: *Slot tagging performance (without LMs) on the MEDIA test set for different training chains for speaker independent and two types of speaker adapted SF models.*

Results with speaker adaptation in terms of CER are shown in Figure 2 for different transfer learning chains. We can see that most of the models with speaker adaptive training (SAT) show better results than speaker independent (SI) ones. SAT models with the proposed zero pseudo i-vector pretraining outperform all SI models and all SAT models obtained with conventional training using i-vectors. In average, the gain from adaptation is greater for less accurate SI models. Table 3 shows the detailed results for different SAT models and their relative comparison with the SI ones.

5. Conclusions

In this paper, we have investigated the effectiveness of speaker adaptation and various transfer learning approaches for end-to-end SLU in the context of the SF task. First, in order to improve the quality of the SF models, during the training, we proposed to use knowledge transfer from an ASR system in another language and from a NER in a target language. Experiments on the French MEDIA test corpus demonstrated that using knowledge transfer from the ASR in English improves the SF model performance by about 14–16% of relative CER reduction for SI models and by 10–20% for speaker adapted models. This approach can be especially useful in a low-resource scenario, when there is a lack of transcribed and semantically annotated data in the target language. The improvement from the transfer learning is greater when the ASR model is trained on the target language (27–37% of relative CER reduction) or when the NER model in the target language is used for pretraining (24–38% of relative CER reduction). Another contribution concerns SAT training for SLU models. We demonstrated that using speaker adaptation can significantly improve the model performance. In addition, for better initialization, we proposed a novel method for SAT, based on zero pseudo i-vector pretraining, which outperforms the conventional SAT models by about $1–21\%$ of relative CER reduction for different models, and SI models – by $5–21\%$. The best adapted system outperforms the best (to our knowledge) published result for this task (for the traditional SLU system) by 17.6% of relative CER reduction.

6. Acknowledgements

This work was supported by the French ANR Agency through the ONTRAC project, under the contract number ANR-18-CE23-0021-01, and by the RFI Atlanstic2020 RAPACE project.

7. References

[1] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[2] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[3] Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[4] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *ASRU*, 2015, pp. 167–174.

[5] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[6] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[7] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *ICASSP*, 2016.

[8] Y. Qian, R. Ubale, V. Ramanaryanan, P. Lange, D. Suendermann-Oeft, K. Evanini, and E. Tsuprun, "Exploring ASR-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system," in *ASRU*, 2017, pp. 569–576.

[9] P. Haghani, A. Narayanan, M. Bacchiani, G. Chuang, N. Gaur, P. Moreno, R. Prabhavalkar, Z. Qu, and A. Waters, "From audio to semantics: Approaches to end-to-end spoken language understanding," *arXiv preprint arXiv:1809.09190*, 2018.

[10] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," *arXiv preprint arXiv:1802.08395*, 2018.

[11] S. Ghannay, A. Caubrière, Y. Estève, N. Camelin, E. Simonnet, A. Laurent, and E. Morin, "End-to-end named entity and semantic concept extraction from speech," in *SLT*, 2018, pp. 692–699.

[12] Y.-P. Chen, R. Price, and S. Bangalore, "Spoken language understanding without speech recognition," in *ICASSP*, 2018.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[15] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *ICASSP*. IEEE, 2016, pp. 4945–4949.

[16] N. Tomashenko and Y. Estève, "Evaluation of feature-space speaker adaptation for end-to-end acoustic models," in *LREC*, 2018.

[17] M. Delcroix, S. Watanabe, A. Ogawa, S. Karita, and T. Nakatani, "Auxiliary feature based adaptation of end-to-end asr systems," *Interspeech*, 2018.

[18] T. Ochiai, S. Watanabe, S. Katagiri, T. Hori, and J. Hershey, "Speaker adaptation for multichannel end-to-end speech recognition," in *ICASSP*. IEEE, 2018, pp. 6707–6711.

[19] K. Li, J. Li, Y. Zhao, K. Kumar, and Y. Gong, "Speaker adaptation for end-to-end CTC models," in *SLT*, 2018, pp. 542–549.

[20] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 55–59.

[21] N. Dehak, P. J. Kenny *et al.*, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 788–798, 2011.

[22] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, pp. 75–98, 1998.

[23] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE transactions on speech and audio processing*, 1994.

[24] N. Tomashenko and Y. Khokhlov, "Speaker adaptation of context dependent deep neural networks based on map-adaptation and gmm-derived feature processing," in *Interspeech*, 2014.

[25] M. Delcroix, K. Kinoshita, A. Ogawa, T. Yoshioka, D. T. Tran, and T. Nakatani, "Context adaptive neural network for rapid adaptation of deep cnn based acoustic models." in *Interspeech*, 2016, pp. 1573–1577.

[26] S. Deena, R. W. Ng, P. Madhyashta, L. Specia, and T. Hain, "Semi-supervised adaptation of rnnlms by fine-tuning with domain-specific auxiliary features," in *Interspeech*. ISCA, 2017, pp. 2715–2719.

[27] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, 2010.

[28] A. Caubrière, N. Tomashenko *et al.*, "Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability," in *Accepted to Interspeech*, 2019.

[29] Y. Esteve, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas, "The EPAC corpus: Manual and automatic annotations of conversational speech in french broadcast news." in *LREC*, 2010.

[30] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of french radio broadcasts," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[31] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the french language," in *LREC*, 2012.

[32] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The REPERE corpus: a multimodal corpus for person recognition." in *LREC*, 2012, pp. 1102–1107.

[33] F. Bechet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, and E. Arbillot, "DECODA: a call-centre human-human spoken conversation corpus." in *LREC*, 2012.

[34] L. Devillers, H. Maynard, S. Rosset, P. Paroubek, K. McTait, D. Mostefa, K. Choukri, L. Charnay, C. Bousquet, N. Vigouroux *et al.*, "The french MEDIA/EVALDA project: the evaluation of the understanding capability of spoken language dialogue systems." in *LREC*, 2004.

[35] F. Lefèvre, D. Mostefa, L. Besacier, Y. Estève, M. Quignard, N. Camelin, B. Favre, B. Jabaian, and L. Rojas-Barahona, "Robustness and portability of spoken language understanding systems among languages and domains: the PortMedia project [in French]," in *JEP-TALN-RECITAL*, 2012, pp. 779–786.

[36] E. Simonnet, S. Ghannay, N. Camelin, and Y. Estève, "Simulating asr errors for training SLU systems," in *LREC 2018*, 2018.

[37] H. Bonneau-Maynard, C. Ayache, F. Bechet, A. Denis, A. Kuhn, F. Lefèvre, D. Mostefa, M. Quignard, S. Rosset, C. Servan *et al.*, "Results of the French Evalda-Media evaluation campaign for literal understanding," in *LREC*, 2006.

[38] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the TED-LIUM corpus with selected data for language modeling and more ted talks." in *LREC*, 2014, pp. 3935–3939.

[39] C. Grouin, S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert, and L. Quintard, "Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview," in *Proceedings of the 5th Linguistic Annotation Workshop*, 2011, pp. 92–100.

[40] V. Vukotic, C. Raymond, and G. Gravier, "Is it time to switch to word embedding and recurrent neural networks for spoken language understanding?" in *Interspeech*, 2015.

[41] C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio, "Batch normalized recurrent neural networks," in *ICASSP*, 2016.

[42] D. Povey, A. Ghoshal *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.