IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Exploring E2E speech recognition systems for new languages

Conrad Bernath[1], *Aitor Álvarez*[1], *Haritz Arzelus*[1], *Carlos-D. Martínez-Hinarejos*[2]

[1]Human Speech and Language Technology Group,
Vicomtech, Spain
[2]Pattern Recognition and Human Language Technologies Research Center,
Universitat Politècnica de València, Spain

[cbernath,aalvarez,harzelus]@vicomtech.org, cmartine@dsic.upv.es

Abstract

Over the last few years, advances in both machine learning algorithms and computer hardware have led to significant improvements in speech recognition technology, mainly through the use of Deep Learning paradigms. As it was amply demonstrated in different studies, Deep Neural Networks (DNNs) have already outperformed traditional Gaussian Mixture Models (GMMs) at acoustic modeling in combination with Hidden Markov Models (HMMs). More recently, new attempts have focused on building end-to-end (E2E) speech recognition architectures, especially in languages with many resources like English and Chinese, with the aim of overcoming the performance of LSTM-HMM and more conventional systems.

The aim of this work is first to present the different techniques that have been applied to enhance state-of-the-art E2E systems for American English using publicly available datasets. Secondly, we describe the construction of E2E systems for Spanish and Basque, and explain the strategies applied to overcome the problem of the limited availability of training data, especially for Basque as a low-resource language. At the evaluation phase, the three E2E systems are also compared with LSTM-HMM based recognition engines built and tested with the same datasets.

Index Terms: speech recognition, deep learning, end-to-end speech recognition, recurrent neural networks

1. Introduction

Automatic Speech Recognition (ASR) systems have historically employed Hidden Markov Models (HMMs) to capture the time variability of the speech signal and Gaussian Mixture Models (GMMs) to model the HMM state probability distributions. Relevant advances in both machine learning and computational capacity have led to significant improvements in the field, mainly by means of Deep Learning algorithms. Thereby, numerous works have shown that Deep Neural Networks (DNNs) in combination with HMMs can outperform GMM-HMM systems at acoustic modeling on a variety of datasets [1].

Recently, new attempts have focused on building End-to-End (E2E) ASR architectures, especially in languages with many resources like English and Chinese [2]. These new architectures intend to benefit from the vast amount of new data available, and to make training a much more efficient process by allowing a better optimization of the system as a whole unit [3]. Besides, one of their most remarkable advantages is to build ASR engines without the need of a phoneme set or a pronunciation lexicon of the language. This definitely reduces the human effort to manually construct resources to help the system to estimate their components. This novel approach allows the system to extract the best features and to employ a single optimization criterion that leads to a better general performance.

The interest in building E2E ASR systems that directly map the input speech signal to grapheme/character sequences and that jointly train the acoustic, pronunciation and language models has exponentially grown in the last years [3, 4, 5, 6].

Two main approaches have been mainly employed to build E2E ASR models. The Connectionist Temporal Classification (CTC) is probably the most widely used criterion for systems based on characters [2, 7, 8], sub-words [9] or words [10]. It employs Markov assumptions and dynamic programming to efficiently solve sequential problems [2, 3, 7]. On the other hand, attention-based methods use an attention mechanism to perform alignment between acoustic frames and characters [4, 5, 6]. Unlike CTC, it does not require several conditional independence assumptions to obtain the label sequence probabilities, allowing extremely non-sequential alignments like in the case of machine translation.

Several techniques have been applied in the literature to enhance the performance of E2E ASR engines. These techniques are mainly focused on compensating for the lack of training data, adapting the characteristics of the system to some specific domains or on gaining robustness in different acoustic environments. *Data augmentation* aims to extend the training material by generating new synthetic data through noise injection or by augmenting the audios using Vocal Tract Length Perturbation (VTLP), tempo modification or speed alteration [11]. The *Transfer Learning* technique is the improvement of a new learning by taking advantage of the knowledge obtained from a previously learned related task [12]. For E2E ASR models, this technique consists of using the weights of a previously trained model in a particular language as initial weights for the training of a new target language. *Fine Tuning* is commonly applied when existing E2E models have to be adapted to some particular conditions. It can be seen as a subtype of *Transfer Learning* but without freezing layers or making changes over them.

A number of enhancement techniques have also been employed to improve the performance of these systems under acoustically noisy conditions. In addition to the front-end methods applied at feature level [13, 14] or training on noisy data with a range of SNR values [15], other methods such as *Dropout* [16] or *Curriculum Learning* [17] have also been proven to improve robustness against noise.

In this work, the construction and evaluation of several E2E systems are presented for English, Spanish and Basque, considering the latter a low resource language. In addition, several new modeling techniques explained above were applied with the aim of outperforming baseline E2E engines and exploring further possibilities within more challenging acoustic domains

10.21437/IberSPEECH.2018-22

and languages. Finally, the results achieved were compared with those obtained by LSTM-HMM based systems through the use of the same training and evaluation datasets.

This paper is organized as follows. Section 2 describes the main architecture of the E2E systems developed within this work. Section 3 describes the training and evaluation data employed, whilst Section 4 presents the baseline ASR systems built for each language including both E2E and LSTM-HMM architectures. The experiments performed are described in Section 5 and finally, Section 6 draws conclusions and looks at the future work.

2. System overview

All the E2E systems presented in this work were developed following the Deep Speech 2 architecture [2]. The core of the system is basically an RNN model, in which speech spectrograms are ingested and text transcriptions are provided as output. Although the Long-short-term Memory (LSTM) is widely used as RNN model, in this architecture Gated Recurrent Units (GRU) [18] are employed, since they have been proven to be trained more rapidly and to be less likely to diverge [19].

A sequence of 2 layers of 2D convolutional neural networks (CNN) are employed as spectral feature extractor from spectrograms. The first layer was composed of 1 input and 32 output channels and it uses filters of size 41×11 and stride of size 2×2. The second layer takes as input the output of the first layer, composed of 32 channels. The output of the second layer incorporates 32 channels as well. This second layer employs a filter dimension of 21×11 and stride of size 2×1. A 2D batch normalization function is applied to the output of both layers, in addition to a *hard tanh* function as an activation function.

The E2E systems are set up using 5 layers of bidirectional GRU layers. Each hidden layer is composed of 800 hidden units. After the bidirectional recurrent layers, a fully connected layer is applied as the last layer of the whole model. The output corresponds to a *softmax* function which computes a probability distribution over characters. This distribution is computed over each timestep. The size of this output layer was equal to the total number of the characters to predict.

During the training process, the CTC loss function is computed to measure the error of the predictions, whilst the gradient is computed using backpropagation through time algorithm with the aim of updating the network parameters. The optimizer is the Stochastic Gradient Descent (SGD).

In addition, external language models (LMs) were integrated during the decoding of the E2E systems with the aim of overcoming the initial results. To this end, modified Kneser-Ney smoothed n-grams models of several orders were estimated using the KenLM toolkit [20].

3. Corpora description

In this section, the acoustic and text data used to train and evaluate both E2E and LSTM-HMM systems are presented for each language.

3.1. English

The freely available corpus LibriSpeech [21], a read speech corpus based on audio-books from LibriVox[1], was used as dataset. The training, development and testing subsets were maintained as original. These partitions are detailed in Table 1.

[1]https://librivox.org/

Table 1: *Training, development and test subsets for English.*

subset	hours
train-clean	464.2
train-other	496.7
dev-clean	5.4
test-clean	5.4
dev-other	5.3
test-other	4.1
test-noisy	5.4

The *clean* partitions correspond to those pools with a lower WER in the whole corpus, whilst the *other* ones contain the most difficult audios *a priori*. *Test-noisy* was artificially created within this work using synthesis by superposition [22] of noise samples to the *test-clean* subset.The noise samples correspond to audios from different acoustic environments selected from the Youtube platform.

Regarding the text data used to train the LMs, they were composed by 22,000 books from Project Gutenberg[2] repository, toting up 803 million tokens and 900,000 unique words.

3.2. Spanish

The Spanish subset of the SAVAS corpus [23] was used as the main dataset. It is composed of broadcast news contents from the Basque Country's public broadcast corporation EiTB (Euskal Irrati Telebista), and includes audios in both clear (studio) and noisy (outside) conditions. This *media* dataset was then transferred through both land- and mobile-lines using different combinations, generating new *telephone* domain subsets, as it is summarized in Table 2.

Table 2: *Training, development and test subsets for Spanish.*

subset	hours
train-media	132.5
train-land-mobile	397.5
dev-media	4
test-media-clean	4
test-media-noisy	4
dev-land	4
test-land	4.6
dev-mobile	4
test-mobile	4.6

Concerning text data, they were obtained by merging transcriptions of the training audios and generic domain news crawled from the Internet. The number of texts summed up a total of 320 million words.

3.3. Basque

The Basque training data was also composed of the Basque subset of the SAVAS corpus, and the audios were gathered from Basque broadcast news programs as well. No *telephone* domain partition was generated in this case. Table 3 describes the main characteristics of the SAVAS Basque corpus.

The text data were also obtained by merging transcriptions and generic news crawled from the Internet. In total, 180 million words were employed for the LMs estimation.

[2]https://www.gutenberg.org/

Table 3: *Training, development and test subsets for Basque.*

subset	hours
train-media	142.25
dev-media	4
test-media-clean	4
test-media-noisy	4

4. Baseline E2E and LSTM-HMM systems

Using the above described corpora, the E2E and LSTM-HMM based baseline systems were trained in order to be compared to the evolved ASR systems presented in Section 5. The characteristics of these baseline systems are described in the following subsections.

4.1. Baseline E2E models

Each E2E model per language was trained over the corresponding corpus described above, through a default setup and without applying any enhancement technique. Linear-scale based spectrograms were employed as input. English E2E baseline models were trained using the train-clean and train-other partitions for 15 epochs and a batch size of 10, whilst the Spanish and Basque baselines were built with the train-media partition for 25 epochs and a batch size of 20 because of the corpus characteristics.

4.2. LSTM-HMM models

These models were estimated with the Kaldi toolkit [24]. The acoustic models corresponded to a hybrid LSTM-HMM implementation, where unidirectional LSTMs were trained to provide posterior probability estimates for the HMM states. Besides, modified Kneser-Ney smoothed 3-gram and 5-gram models were used for decoding and re-scoring of the lattices respectively. Both LMs were estimated with the KenLM toolkit.

5. Experiments

5.1. Linear- and Mel-scale based spectrograms

The original Deep Speech 2 architecture employs linear spectrograms as the main audio parametrization method. This experiment was focused on analyzing the use of Mel-scale based spectrogram as input data to the E2E model as it was employed in [25], as it aims to highlight the most relevant information for the human hearing according to the Mel scale.

Two models were compared for the three languages under evaluation; the above presented baseline E2E models trained using linear spectrograms, and the evolved new models estimated with the same configuration and training-set but using Mel-based spectrograms for audio parametrization.

The results in Table 4 show a noticeable improvement for all the experiments except one when using Mel-scale based spectrograms. Focusing on the results without using LM (No LM), a relative improvement of 3% was reached for English on the clean test set, whilst these enhancements were of 8.7% and 10% for Spanish and Basque. This positive behavior was also maintained over the noisy test set, obtaining relative improvements of 1.4%, 7.0% and 6.7% for each corresponding language. The results using LM followed the same tendency, with remarkable differences obtained specially for Basque over the test-media-clean set, where a real improvement of 4 percentage points was reached (from 12.9 to 8.9).

Table 4: *WER (%) results for linear- and Mel-based systems for each language.*

	Test(-media-)clean		Test(-media-)noisy	
	No LM	**3-gram**	**No LM**	**3-gram**
English				
Baseline-EN	10.6	5.6	35.3	23.6
Mel-scale-EN	10.2	5.4	34.8	22.2
Spanish				
Baseline-ES	24.0	10.3	39.5	19.2
Mel-scale-ES	21.9	10.3	36.7	18.9
Basque				
Baseline-EU	23.8	12.9	38.8	17.3
Mel-scale-EU	21.2	8.9	36.2	19.2

5.2. Training a mixed model for media and telephone

Telephone speech sampling is usually limited by the bandwidth at which the audio is transmitted through the telephone channel, getting values as much at 8 kHz (4 kHz Nyquist). This commonly implies to train separate acoustic models per domain.

In this work, a mixed E2E model for the media and telephone domains was built, and then a fine-tuning technique was applied over each domain training set to generate domain adapted individual models. It was only performed for the Spanish language and the performance of the resulting systems was compared to the Spanish baseline system. Following the results obtained in the previous experiment, and with the aim of loosing the less valuable telephone speech information as possible, Mel-scale based spectrograms were employed for audio parametrization. In addition, the Spanish train-media dataset was 3-fold augmented through speed perturbation to balance the quantity of data in both domains.

The results achieved over the *media* test set and the *telephone* test set are presented in Tables 5 and 6 respectively.

Table 5: *WER results for the baseline, mixed (Mixed-ES) and fine-tuned (Mixed-FT-Media and Mixed-FT-Phone) models over the Spanish media test set.*

	Test-media-clean		Test-media-noisy	
	No LM	**3-gram**	**No LM**	**3-gram**
Baseline-ES	24.0	10.3	39.5	19.2
Mixed-ES	16.5	8.8	21.3	11.5
Mixed-FT-Media	15.7	8.5	20.5	10.9
Mixed-FT-Phone	18.6	9.6	22.4	11.5

Table 6: *WER results for the baseline, mixed (Mixed-ES) and fine-tuned (Mixed-FT-Media and Mixed-FT-Phone) models over the Spanish telephone test set.*

	Test-mobile		Test-land	
	No LM	**3-gram**	**No LM**	**3-gram**
Baseline-ES	59.5	33.8	51.7	25.9
Mixed-ES	22.4	12.8	18.4	9.9
Mixed-FT-Media	23.4	13.0	18.7	10.0
Mixed-FT-Phone	22.2	12.1	17.2	9.3

Looking at the results obtained for the *media* test set in clean conditions, relative improvements of 7.5% and 8.3% of the Mixed-ES and the fine-tuned Mixed-FT-Media models can be observed with respect to the baseline. Besides, improvements of 18.3% were achieved by the fine-tuned Mixed-FT-Media model for the *media* noisy test set when comparing to the baseline system.

Regarding the results on the *telephone* domain presented in Table 6, the three models under evaluation present better performance than the Baseline-ES model. As it was expected, the Mixed-FT-Phone model shows the best results with a real improvement of 37.3% with respect to the baseline over the Test-mobile set, composed of audios transferred by the mobile channel. This improvement achieves a 34.5% in the case of the Test-land set. The Mixed-ES model also shows notable enhancements, with slightly worse WER than the Mixed-FT-Phone model. The Mixed-FT-Media performs satisfactorily as well, but it shows a little lower performance than the others.

Finally, it can be also observed that the Mixed-ES model performs almost as well as the fine-tuned models even without having applied any fine-tuning technique.

5.3. Best E2E models compared to LSTM-HMM systems

The aim of this task was to select the best E2E models presented above and compare them to (1) the state-of-the-art models in the literature if available and (2) LSTM-HMM based systems developed on top of the Kaldi toolkit. For this experiment, the selected models for each language were called as Evolved-EN, Evolved-ES and Evolved-EU.

- Evolved-EN. It corresponded to the English baseline model evolved through fine-tuning and a speed-based data augmentation techniques applied for 10 more epochs. The decoding was performed using an external 3-gram LM and using a beam size of 1000.
- Evolved-ES. This model refers to the Spanish fine-tuned Mixed-FT-Media. It was decoded with two configurations; (1) using an external 3-gram LM and a beam size of 600 and (2) with a 5-gram LM with a beam size of 1000.
- Evolved-EU. The selected Basque model was the Mel-scale-EU model presented in subsection 5.1. It was decoded using external 3-gram and 5-gram LM models with a beam size of 600 and 1000 respectively.

In Table 7, a comparison between the evolved model, the Kaldi based LSTM-HMM model, and the results obtained from reference systems in the literature are presented for the English language. Since Test-noisy was created for this work, the results for this subset are only given for the first two models.

Table 7: *WER comparison between our best model, the LSTM-HMM model and reference systems for English language.*

	Test-clean	**Test-other**	**Test-noisy**
Evolved-EN	4.9	15.4	21.0
LSTM-HMM	6.0	15.2	26.3
DeepSpeech-1 [2]	7.8	21.7	-
DeepSpeech-2 [2]	5.3	13.2	-
Human [2]	5.8	12.6	-
Wav2Letter [8]	6.9	-	-

Table 8: *WER comparison of Kaldi vs E2E for SAVAS Spanish.*

	Test clean		**Test noisy**	
	3-gram	**5-gram**	**3-gram**	**5-gram**
Spanish				
Evolved-ES	8.5	7.2	10.9	9.3
LSTM-HMM	7.9	7.7	11.9	10.8
Basque				
Evolved-EU	8.9	6.6	19.2	15.9
LSTM-HMM	7.8	6.0	10.8	8.2

As it can be observed in Table 7, the evolved E2E system outperformed the LSTM-HMM system and all the reference systems over the clean test. It shows the effect of applying Mel-scale based parametrization, and techniques like fine-tuning and speed-based data augmentation. It has to be also remarked that the WER obtained is 0.9 percentage points lower than the achieved by a human manual transcription.

On the contrary, for the Test-other partition, the LSTM-HMM model shows a better performance presenting an error rate 0.2 percentage points lower than the proposed E2E model. With respect to the reference systems, the Evolved-EN model obtains a better performance than Deep Speech 1 model, but still, an error of 2.2 points higher than Deep Speech 2. It has to be considered that the model from Deep Speech 2 was trained with 12,000 hours (including the Librispeech corpus), more than ten times more than the Evolved-EN system (960h).

Table 8 shows how the best Spanish E2E model overcame all the results obtained by the LSTM-HMM based system with the exception of the Test clean when a 3-gram was applied. It shows the robustness of the fine-tuned mixed model trained with data from different acoustic domains. Finally, the scarcity of training data for Basque benefited the LSTM-HMM model against the E2E model, which was estimated using Mel spectrograms as the only enhancement technique.

6. Conclusions and Future Work

In this work, E2E ASR systems for English, Spanish and Basque have been developed and evaluated against reference and LSTM-HMM architectures. Besides, the positive impact of applying some enhancement techniques have been demonstrated through different experimental evaluations.

As main conclusions, it can be stated that using Mel-scale based spectrograms overcomes linear-based ones, as it was proven in all the experiments. Moreover, it was shown that robust hybrid E2E models that perform almost as well as in-domain models can be generated for acoustically different environments. With regard to the training data, as in the case of Basque compared to English, it was clear that more data leads to better results. However, when the resources are limited, using an external LM of a high order can improve the performance, as it was shown for all the languages under evaluation.

As a future work, one important task will be the generation of new Spanish and Basque E2E models with more training data, since these models were still weak without LM. Moreover, current N-grams will be replaced by RNN based LMs, specially for Basque as an agglutinative language. Finally, following novel studies, new research efforts will be made to develop methodologies for a semi-supervised learning within E2E architectures.

7. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II–1764–II–1772.

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

[5] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[6] L. Lu, X. Zhang, and S. Renals, "On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5060–5064, 2016.

[7] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 167–174.

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

[9] H. Liu, Z. Zhu, X. Li, and S. Satheesh, "Gram-ctc: Automatic unit selection and target decomposition for sequence labelling," *arXiv preprint arXiv:1703.00096*, 2017.

[10] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, and D. Nahamoo, "Direct acoustics-to-word models for english conversational speech recognition," *arXiv preprint arXiv:1703.07754*, 2017.

[11] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 2010, pp. 242–264.

[13] Y. Fujita, R. Takashima, T. Homma, R. Ikeshita, Y. Kawaguchi, T. Sumiyoshi, T. Endo, and M. Togami, "Unified asr system using lgm-based source separation, noise-robust feature extraction, and word hypothesis selection," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 416–422.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[15] D. Palaz, R. Collobert *et al.*, "Analysis of cnn-based speech recognition system using raw speech as input," in *Proceedings of INTERSPEECH*, no. EPFL-CONF-210029, 2015.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] S. Braun, D. Neil, and S.-C. Liu, "A curriculum learning method for improved noise robustness in automatic speech recognition," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 548–552.

[18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[19] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *International Conference on Machine Learning*, 2015, pp. 2342–2350.

[20] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011, pp. 187–197.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[22] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[23] A. del Pozo, C. Aliprandi, A. Álvarez, C. Mendes, J. P. Neto, S. Paulo, N. Piccinini, and M. Raffaelli, "Savas: Collecting, annotating and sharing audiovisual language resources for automatic subtitling." in *LREC*, 2014, pp. 432–436.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[25] V. Liptchinsky, G. Synnaeve, and R. Collobert, "Letter-based speech recognition with gated convnets," *CoRR*, vol. abs/1712.09444, 2017. [Online]. Available: http://arxiv.org/abs/1712.09444