Confidence measure for speech-to-concept end-to-end spoken language understanding

Antoine Caubrière[1], *Yannick Estève*[2], *Antoine Laurent*[1], *Emmanuel Morin*[3]

[1]LIUM - Le Mans University, France
[2]LIA - Avignon University, France
[3]LS2N - Nantes University, France

antoine.caubriere@univ-lemans.fr, yannick.esteve@univ-avignon.fr,
antoine.laurent@univ-lemans.fr, emmanuel.morin@univ-nantes.fr

Abstract

Recent studies have led to the introduction of Speech-to-Concept End-to-End (E2E) neural architectures for Spoken Language Understanding (SLU) that reach state of the art performance. In this work, we propose a way to compute confidence measures on semantic concepts recognized by a Speech-to-Text E2E SLU system. We investigate the use of the hidden representations of our CTC-based SLU system to train an external simple classifier. We experiment two kinds of external simple classifiers to analyze subsequences of hidden representations involved in recognized semantic concepts. The first external classifier is based on a MLP while the second one is based on a bLSTM neural network. We compare them to a baseline confidence measure computed directly from the softmax outputs of the E2E system. On the French challenging MEDIA corpus, when the confidence measure is used to reject, experiments show that using an external BLSTM significantly outperforms the other approaches in terms of precision/recall. To evaluate the additional information provided by this confidence measure, we compute the value of Normalised Cross-Entropy (NCE). Reaching a value equal to 0.288, we show that our best proposed confidence measure brings relevant information about the reliability of a recognized concept.

Index Terms: Confidence Measure, End-to-End, Spoken Language Understanding, Neural networks

1. Introduction

Speech-to-concept end-to-end (E2E) neural architectures for Spoken Language Understanding (SLU) have been recently introduced [1, 2, 3, 4, 5] and reach state of the art performance on well known benchmark datasets for intent recognition and slot filling task [6]. Usually, such tasks are related to human-machine dialogue applications and consist in extracting values of semantic concepts expected by the dialogue manager. Until two years ago, SLU systems working on slot filling task were based on a treatment chain, first composed of an automatic speech recognition (ASR) system, often followed by different kinds of natural language processing (NLP) like part of speech tagging applied on ASR output, and last the natural language understanding (NLU) applied to the enriched ASR system in order to extract semantic concepts and their values. Thanks to speech-to-concept end-to-end neural architectures, values and semantic concepts can now be directly extracted from speech by using a single deep neural model, sometimes used in conjunction with beam-search decoding that refines the neural output through a language model rescoring.

When a concept/value pair is recognized, this pair is provided to a dialogue manager that handles the spoken human-machine interactions. Despite the recent advances in this topic, SLU systems continue to make errors, particularly on complex tasks for which the internal semantic representation is challenging, due to the difficulty of the dialogue objectives [7]. To make possible a better dialogue management able to handle SLU potential errors, a solution is to provide a confidence measure given by the SLU system to the dialogue manager, for each recognized concept/value pair [8, 9, 10, 11, 12]. A such confidence measure can be used to reject some recognized concept/value pairs with low confidence value. But a such confidence measure can also be used in a more subtle way, for instance by permitting the dialogue manager to apply a strategy of implicit confirmation on low reliable recognized concepts without wasting time by asking for an explicit confirmation (*e.g.* the system can ask: *"For your stay on October 28th, do you want a room with twin beds?"*, while the concept/value *date[October 28]* recognized in the previous user turn was not reliable according to its confidence measure). Approaches to compute such confidence measures have been proposed by the past on SLU modules based on treatment chain (ASR+ NLP + NLU) [8, 9, 11].

On our knowledge, this paper is the first study that investigates confidence measures in the framework of speech-to-concept end-to-end neural architecture for SLU. While confidence measures for classical treatment chain are usually based on the combination of acoustic, linguistic, semantic and/or decoder information[10], we suggest in our approach to take benefit from the analysis of the hidden representations computed during inference.

After a description of our E2E SLU system and the MEDIA dataset (section 2 and section 3), we present in section 4 a method to compute vectorial representations of concept/value pairs from hidden representations extracted from the E2E SLU model. Then we describe the proposed confident measure (section 5) and present experimental results (section 6).

2. SLU system description

In this work, we use the E2E SLU system we introduced in previous works [6, 2, 13]. Similar to Deep Speech 2 [14], its architecture consists of a stack of two 2D-invariant convolutional layers (CNN), five bidirectional long short term memory layers (bLSTM) with sequence-wise batch normalization and a softmax layer.

This system is trained with the Connectionist Temporal Classification (CTC) loss function [15]. This function allows

This work was supported by the RFI Atlanstic2020 RAPACEproject and the AISSPER project through the French NationalResearch Agency (ANR) under Contract AAPG 2019 ANR-19-CE23-0004-01.

the system to learn an alignment between an audio input and a character sequence to produce. Input features are sequences of log-spectrograms of power normalized audio clips calculated on 20ms windows. Output sequences consist of a character sequence composed of word and semantics concepts. Semantics concepts are represented by starting and ending tags before and after the words supporting these concepts. Starting tags defines the nature of concepts while ending tag will only close an opened tag. The number of starting tags depends on the targeted task, and we use only one ending tag. In this way, an example of an output sequences of this system could be "I would like <**nb_room** two > <**room_type** double-bed rooms >". In this sequence, <**nb_room** and <**room_type** are two starting tags defining respectively the semantics concepts "number of room" and "room type", while the '>' symbol represents the unique ending tags. Since our system provides character-based outputs, starting and ending tags are represented by a single character within the sequence to be produced by the neural network. Previous example become "I would like **ł** two > **ø** double-bed rooms >", where '**ł**' is "<**nb_room**" and '**ø**' is "<**room_type**".

Our end-to-end SLU system is trained following the Curriculum-based Transfer Learning approach (CTL) we proposed in [6]. It consists to train the same model through a sequence of training processes and transfer learning. This process sequence follows a curriculum strategy for which tasks are ordered from the most generic one to the most specific one. We use 3 kinds of tasks: first speech recognition (ASR), then named entity recognition (NER), and finally semantic concept extraction (SLU). We define this order of specificity because of the lack of semantic concepts for the speech recognition task and the more generic nature of named entities in front of the semantic concepts. Named entity recognition task is trained following the same way as the semantic concept extraction task. We add boundaries of named entity concepts inside the character sequences to be produced and boundaries are encoded into single characters. Transfer learning is applied between each task and we keep all the parameters of the produced model of a current training step as initialization of the next training step, except the softmax layer. Parameters of this top layer are fully reset because of the change of possible output labels at each training step. Thanks to this strategy, our end-to-end models reached state-of-the-art performance. More details are given in [6].

3. The MEDIA benchmark dataset

The MEDIA corpus is a French dataset of audio recordings with manual annotations, dedicated to semantic extraction from speech in a context of human/machine dialogues. The corpus has manual transcriptions and semantic annotations of dialogues from 250 speakers. It is split into the following three parts [16]: (1) the training set (720 dialogues, 12K sentences, 31.7K semantic concepts), (2) the development set (79 dialogues, 1.3K sentences, 3.3K semantic concepts), and (3) the test set (200 dialogues, 3K sentences, 8.8K semantic concepts). A concept is defined by a label and a value, for example the value *2001/02/03* can be associated to the concept *date* [16, 17, 3]. The MEDIA corpus is related to the hotel booking domain, and its annotation contains 76 semantic concept tags: *room number, hotel name, location, date, room equipment*, etc. In [7], the authors show that the MEDIA benchmark dataset is one of the most challenging SLU benchmark available, largely more complex than the widely used ATIS corpus.

4. From hidden representation sequences to concept embeddings

As described in section 2, our E2E SLU system is based on the Deep Speech 2 architecture (mainly composed of CNN and RNN layers). It is trained with the CTC loss function that permits the neural model to provide character sequences from speech input. The outputs are single symbols composed of alphabet letters, space, apostrophe, the blank symbol (that simulates a void emission in the CTC paradigm), but also single symbols related to semantic concepts as seen above. For each output time-step t, the RNN makes a prediction, $p(l_t|x)$, where l_t is a single symbol. That means that to produce a word, or a concept and its value, several symbols l_i will be successively predicted. For instance, the concept/value pair *room_equipement[twin bed]* can be recognized thanks to the following symbol sequence produced symbol by symbol by the E2E neural model: ρ *tt_wwiii_nn bbeeee_d* $>$, where ρ is a symbol that both indicates the beginning of a concept and categorizes this concept as being *room_equipement*, $>$ is the symbol that indicates the end of the concept, and _ is the blank symbol. In this example, the entire sequence (from ρ to $>$), contains 26 symbols. That means that 26 consecutive predictions have been made, one at each time-step t_k. For each time-step, the E2E model computes hidden representations.

In order to get a vectorial representation of a concept/value pair supported by all the symbols predicted from time-step t_i to time-step t_j, we suggest to extract the hidden representations of the n^{th} RNN layer computed from t_i to t_j, as showed in figure 1. We can use this entire sequence as is, or compute a simple average of all the hidden representations to get a unique vector.

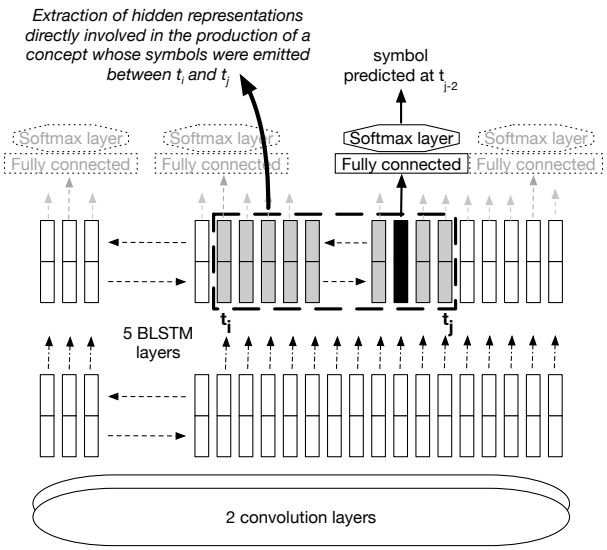

Figure 1: *Extraction of hidden representations of semantic concepts.*

5. Confidence Measures

We investigate the use of its vectorial representations (single vector or vector sequence) extracted from the E2E neural model to compute a confidence measure of a concept/value pair recognized by this model. For that, we feed a simple external classifier with these representations and train the classifier in order to recognize the correct semantic concepts from these represen-

tations. At inference time, we expect to use the output scores of this external classifier to get a confidence measure on the concept/value pair recognized by the E2E model thanks to the hidden representation. The use of an external simple classifier applied to hidden representations is inspired by recent studies like in [18], where the authors analyzed the internal representation of an end-to-end automatic speech recognition system in regards with phonetic information.

The use of two kinds of simple classifiers are investigated:

- a MultiLayer Perceptron (MLP) to analyze the unique vector (800 dimensions) computed by averaging the sequence of hidden representations extracted for a recognized concept/value. This classifier is composed of one hidden layer (200 dimensions, ReLU activation function) and a softmax layer (76 dimensions, equal to the number of semantic concepts in our experiment). This classifier was trained by using the Adam optimizer [19] with the recommanded parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$), except the learning rate which is set to 10^{-4}. We use the cross-entropy loss function and a batch size of 20.
- Or a bLSTM neural network that is fed by the sequence of hidden representations extracted for the recognized concept/value. This classifier is composed of one bidirectional LSTM layer of 200 units followed by a softmax layer dimension. We train this classifier in the same way as the MLP classifier.

In order to compare the quality of the single averaged vector and the sequence of hidden representations, we carried out experiments on semantic concept recognition from these vectorial representations. The averaged vector is used as input of the MLP classifier, while the hidden representation sequence is used as input of the bLSTM classifier. To train the classifiers, we extract E2E internal representations computed on training and validation dataset. This extraction was also made from different layers to compare the evolution of the hidden representations in function of the layer level, in regards with their capability to bring relevant semantic information. For these experiments, we take into account only the representation of the well-recognized concept/value pairs. Table 1 presents the results obtained from both classifiers when hidden representations are extracted from different levels of hidden layers (from the first one to the fifth one).

Table 1: *Comparison of vectorial representations and their relevant classifier in terms of accuracy on internal representations of well-recognized concepts on the MEDIA validaton dataset*

targeted E2E BLSTM layer	unique vector MLP accuracy	vector sequence bLSTM accuracy
1.	70.94%	90.42%
2.	82.15%	94.32%
3.	90.24%	97.14%
4.	**96.87%**	98.73%
5.	95.41%	**99.32%**

For both classifiers, results show an improvement of internal representations throughout the hidden layers of the E2E model. MLP classifier accuracy going stepwise from 70.94 % (1) to 95.41 % (5) with a peak at 96.87 % (4). A similar scenario for the bLSTM classifier can be observed with accuracy going stepwise from 90.42 % (1) to 99.32 % (5). Both classifiers reach good accuracy with the representation from the last hidden layer, especially the bLSTM classifier which is able to reach more than 99 % of accuracy. This can be explained by the rich representation of the sequence of hidden representations, in comparison to their combination through an average than loses dynamics and variability when computing a unique vector.

Subsequently, we propose to consider as a confidence measure the softmax value of a simple external classifier (MLP or bLSTM) for each concept/pair recognized by the E2E model. When a concept/value pair is recognized by the E2E neural model, corresponding hidden representations from the last E2E BLSTM layer are extracted as described in section 4. Then, these hidden representations are presented to an external classifier, either as an averaged unique vector to a MLP, either as a sequence of hidden representations to a BSLTM, as described above. The score provided by the external classifier softmax layer to the concept initially recognized by the E2E neural model is considered as a confidence measure related to this recognition.

6. Experiments

In this section, we investigate the capability of a simple external classifier to provide reliable confidence measures through their softmax outputs scores. Simple classifiers are trained on the concept representations extracted from the last hidden layer of the E2E SLU model, as described in the previous section. On the validation and the test datasets, each recognized concept/value pair is processed through a simple classifier. As a baseline confidence measure, we consider the average of the softmax scores provided by the E2E model for each symbol involved in the symbol sequence that supports a concept/value pair. This baseline measure is called *E2E softmax avg*.

We carried out these experiments on two variants of our E2E SLU system. The first one is the system described in section 2. We called it *normal mode*: it provides both speech transcription and semantic concepts. The second system is similar to the first one and takes benefit from the *star mode* introduced in [1]. It consists of training the E2E system to recognize only concept/value pairs, without recognizing words that are not involved in a concept value. During the training process, all words that are between two concepts are replaced by the CTC blank symbol. This allows the CTC loss function to focus more on concept/value pairs instead of words that do not contain relevant semantic information. Performances of both systems in terms of Concept Error Rate (CER) and Concept Value Error Rate (CVER) are given in table 2. More details and results are given in [6].

Table 2: *Performance measured on Concept Error Rate and Concept/Value Error rate on the MEDIA test dataset for the two variants of the E2E SLU system*

E2E variant	CER	CVER
normal mode	21.6%	27.7%
star mode	20.1%	26.9%

Thanks to confidence measures, it is possible to apply a threshold in order to reject unreliable concept/value pairs. Figure 2 presents the precision/recall curves computed after applying a filtering threshold on confidence measures related to concept/value pairs recognized by the E2E model in *normal mode* on the MEDIA test dataset. Green curve corresponds to the baseline confidence measure, orange to the confidence measure based on the external MLP classifier, and blue to the external bLSTM.

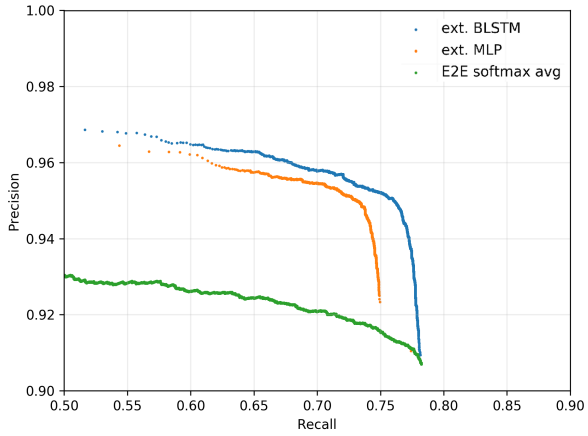

Figure 2: *Precision/Recall after filtering applied to different confidence measures on concept/value pairs recognized by the E2E SLU model on MEDIA test dataset in normal mode*

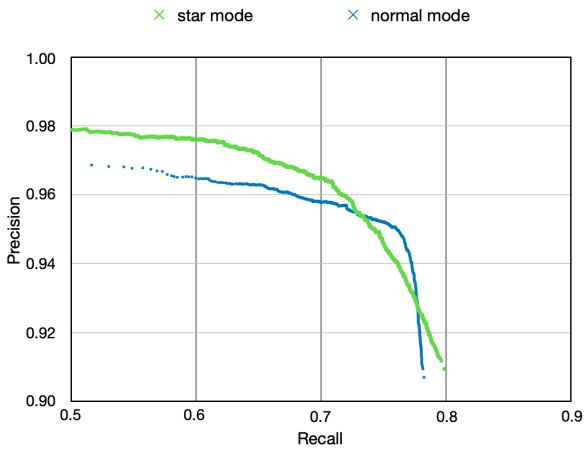

Figure 3: *Precision/Recall after filtering applied to the external bLSTM-based confidence measure on concept/value pairs recognized by each one of the two E2E modes on MEDIA test*

These results show that external classifiers provide a more reliable confidence measure than a baseline measure based on the E2E softmax values. Among the confidence measures based on external classifiers, the one based on the bLSTM gives the best results. This is consistent with the results presented in table 1 that showed that the external bLSTM is more precise to recognize semantic concepts from E2E hidden representations than the MLP. Figure 3 compares results between the external bLSTM-based confidence measures applied to outputs of the E2E model working in *star mode* (green) or *normal mode* (blue). It is interesting to notice that while E2E *star mode* initially provides better results than E2E *normal mode* in terms of CER or CVER without filtering, these results reveal that when the recall is between 0.74 and 0.78 the *star mode* does not outperform the *normal mode*. Last, an interesting point is to be reported. In [12], the authors report the best precision point never reached before in the MEDIA data for concept/value pair recognition: 0.89 of precision, with a recall of 0.68. To reach that, they combine four classical pipelines SLU systems. Our results show that it is possible to get very better results with an end-to-end speech-to-concept model by using the confidence measure proposed in this paper. For instance, for the same recall value of 0.68, we are now able to reach a precision of 0.95 on concept/value recognition.

In addition to this analysis of confidence measure capability to reject unreliable recognized concept/value pairs, it is interesting to evaluate their use in order to predict the probability that recognition is correct. To evaluate this in speech recognition, the normalized cross-entropy (NCE) metric is commonly used [20], for instance in NIST evaluation campaigns. This metric is an information-theoretic measure of how much additional information the confidence measure provides. A positive value of NCE means that the confidence scores provide useful extra information.

The NCE metrics is defined as:

$$NCE = \frac{H_{max} + \sum_{C_{cor}} \log_2(m(C)) + \sum_{C_{uncor}} \log_2(1 - m(C))}{H_{max}} \quad (1)$$

where: $H_{max} = -n \log_2(P) - (N - n) \log_2(1 - P)$, n is the number of concept/value pairs correctly emitted, N is the total number of emitted concept/value pairs, $P = \frac{n}{N}$ is the average probability that a concept/value pair is correctly recognized, and $m(C)$ is the confidence measure of the current concept/value pair. Since it gave the best results on previous experiments, we focus our NCE evaluation on the confidence measure based on an external bLSTM classifier. To calibrate the bLSTM score, we apply to this score a piece-wise linear mapping instead of using it directly as a confidence measure. This kind of mapping has been successfully used for speech recognition for at least 20 years [20]. Calibration is optimized on the validation dataset. The same one is then applied to the test dataset. NCE scores are reported in table 3 for the external bLSTM-based confidence measure computed from either E2E *normal mode* or E2E *star mode*. Notice that the calibration mapping is specific to the E2E mode.

Table 3: *Reliability evaluated in NCE values of the external bLSTM-based confidence measures applied to each of the two E2E modes*

E2E SLU mode	development	test
Normal mode	0.226	0.288
Star mode	0.195	0.241

Results show that mappings defined by observations on the development set still relevant for the test set. By reaching 0.288 on the E2E *normal mode* outputs and 0.241 on the E2E *star mode* outputs on test dataset, there results confirm that the proposed confidence measures provides relevant additional information.

7. Conclusion

This study proposes an approach to compute relevant confidence measures for speech-to-concept end-to-end SLU systems. These confidence measures are based on the use of simple external classifiers that process variable-length hidden representations extracted from the end-to-end SLU model. They permit to label each concept/value pair recognized by the end-to-end SLU model by a confidence measure value. Experiments show that the best proposed confidence measure brings relevant information measured through an NCE score of 0.288. Also, we observed that by applying a filtering threshold to this confidence measure values, it is possible to reach a precision at least equal to 0.94 with a recall of 0.77, that was never reached before on the MEDIA benchmark dataset.

8. References

[1] 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 *IEEE Spoken Language Technology Workshop*, Athens, Greece, 2018, pp. 692–699.

[2] N. Tomashenko, A. Caubrière, and Y. Estève, "Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech," *Interspeech*, 2019.

[3] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[4] P. Wang, L. Wei, Y. Cao, J. Xie, and Z. Nie, "Large-scale unsupervised pre-training for end-to-end spoken language understanding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7999–8003.

[5] R. Price, "End-to-end spoken language understanding without matched language speech model pretraining data," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7979–7983.

[6] A. Caubrière, N. Tomashenko, A. Laurent, E. Morin, N. Camelin, and Y. Estève, "Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability," in *Interspeech 2019*, Graz, Austria, Sep. 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02304597

[7] F. Béchet and C. Raymond, "Benchmarking benchmarks: introducing new automatic indicators for benchmarking spoken language understanding corpora," in *Interspeech 2019*, Graz, Austria, Sep. 2019.

[8] T. J. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," *Computer Speech & Language*, vol. 16, no. 1, pp. 49–67, 2002.

[9] D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tur, "Beyond asr 1-best: Using word confusion networks in spoken language understanding," *Computer Speech & Language*, vol. 20, no. 4, pp. 495–514, 2006.

[10] C. Raymond, Y. Esteve, F. Béchet, R. De Mori, and G. Damnati, "Belief confirmation in spoken dialog systems using confidence measures," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 150–155.

[11] B. Minescu, G. Damnati, F. Béchet, and R. D. Mori, "Conditional use of word lattices, confusion networks and 1-best string hypotheses in a sequential interpretation strategy," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[12] E. Simonnet, S. Ghannay, N. Camelin, Y. Estève, and R. De Mori, "Asr error management for improving spoken language understanding," *arXiv preprint arXiv:1705.09515*, 2017.

[13] A. Caubrière, S. Ghannay, N. Tomashenko, R. De Mori, A. Laurent, E. Morin, and Y. Estève, "Error analysis applied to end-to-end spoken language understanding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8514–8518.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Y. Belinkov and J. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *Advances in Neural Information Processing Systems*, 2017, pp. 2441–2451.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*, vol. 27. Baltimore, 2000, pp. 78–81.