Weakly-supervised text-to-speech alignment confidence measure

Guillaume Serrière, Christophe Cerisara, Dominique Fohr, Odile Mella

Weakly-supervised text-to-speech alignment confidence measure

Guillaume Serrière, Christophe Cerisara, Dominique Fohr, Odile Mella
LORIA URM 7503, 54506 Vandoeuvre-les-Nancy, France
guillaume.serriere@telecomnancy.net
{cerisara,fohr,mella}@loria.fr

Abstract

This work proposes a new confidence measure for evaluating text-to-speech alignment systems outputs, which is a key component for many applications, such as semi-automatic corpus anonymization, lips syncing, film dubbing, corpus preparation for speech synthesis and speech recognition acoustic models training. This confidence measure exploits deep neural networks that are trained on large corpora without direct supervision. It is evaluated on an open-source spontaneous speech corpus and outperforms a confidence score derived from a state-of-the-art text-to-speech aligner. We further show that this confidence measure can be used to fine-tune the output of this aligner and improve the quality of the resulting alignment.

1 Introduction

This work focuses on the text-to-speech alignment (T2SA) task, which consists in temporally aligning a given speech sound file with its known text transcription. The standard objective quality metric is the expected alignment error, measured in seconds and defined as $\mathcal{L} = E[|\hat{t} - t|]$, where t is the gold timestamp of a word boundary, and $\hat{t}$ the corresponding timestamp estimated by the aligner (Keshet et al., 2005).

Text-to-speech alignment is an important task for many applications, including: (i) Lip-syncing in cartoons production and film dubbing; (ii) Anonymization of audio corpus; (iii) Pre-processing of audio corpora for training new speech recognition systems; (iv) Indexing audio-visual corpora for browsing and querying; (v) Sampling sounds for speech synthesis; (vi) Second-language learning.

We propose in this work a novel confidence measure for detecting erroneous word boundaries at the output of an existing T2SA system. Accurately detecting misplaced word boundaries is crucial to reduce post-processing costs in every previous application. For instance, only the most reliable segments may be chosen for acoustic model training and speech synthesis, and manual corrections may be limited to the less reliable boundaries for lip syncing and corpus anonymization.

The proposed confidence measure is computed by a deep neural network (DNN) that is trained on a large corpus without any manually annotated word boundaries. We show on a gold corpus of French spontaneous speech that the proposed model is able to detect correct boundaries with a significantly better accuracy than the acoustic models used in the T2SA system, thanks to the acoustic features automatically captured by the deep neural model on the large unlabelled corpus. We further show that the proposed confidence measure may be used to post-process the T2SA output and improve its precision.

2 Related works

Every text-to-speech aligner faces three main challenges: (i) Handling imperfect transcriptions; (ii) Supporting noisy acoustic conditions; (iii) Finding the globally optimal alignment on long (up to a few hours) audio files. Many solutions have been proposed to address these issues. For instance, "anchor-based" approaches (Moreno et al., 1998; de Jong et al., 2006; Hazen, 2006) automatically infer high-confidence words timestamps at regular intervals in a long audio file in order to enable regular batch

alignment between two successive anchors. The issue of aligning highly imperfect text to speech may be addressed with standard acoustic adaptation (Zhao et al., 2005), or by performing recognition at the phoneme level only with monophthongs and fricatives, which appear to be more robust to noise than other phonemes (Haubold and Kender, 2007). Complementary, better phonetization models of unknown words may also be used (Bigi, 2013). More generally, various models have been proposed for phoneme alignment, such as discriminative shallow large-margin alignment models in (Keshet et al., 2005), non-neural unsupervised acoustic models in (Milde, 2014) and in (Lanchantin et al., 2015), where two DNNs are used respectively for acoustic modelling for speech transcription and for segmenting the speech file into speech and non-speech segments. A remarkable HMM-based architecture is also proposed in (Brognaux and Drugman, 2016), where the acoustic models are trained solely on the target corpus to align. The authors of (Yuan et al., 2013) demonstrate the importance of producing high-precision temporal limits with dedicated models, and propose in (Stolcke et al., 2014) a neural network to fine-tune the alignment. We follow this line of work, but rather focus on estimating the actual quality of the proposed boundaries with a confidence measure. A confidence measure is proposed in the aligner ALISA (Stan et al., 2016), but its role is to filter-out wrongly recognized sentences. Conversely, few publications address the problem of detecting reliable temporal boundaries after T2SA. (Paulo and Oliveira, 2004) proposes a confidence measure that is based on a synthetic speech signal, while (Keshet et al., 2005) discriminatively trains base functions that define an alignment confidence measure, but which is not evaluated per se: thanks to the decomposability property of the base functions, this measure is rather used with a dynamic programming algorithm to output a final alignment. Our work is, to the best of our knowledge, the first proposal to use the modelling potential of deep networks to compute successful confidence measures of text-to-speech alignment outputs.

3 Proposed model

3.1 Model description

The proposed model is shown in Figure 1. Two models, respectively called the *Boundary inspector* and the *Boundary selector*, are built to compute a confidence measure that any candidate word boundary is correct or not. Both the *Boundary inspector* and *selector* take as input an acoustic window of ± 0.05s around the candidate word boundary, plus two categorical inputs representing the left and right phonemes. They can thus be viewed as acoustic models that are specialized in identifying boundaries between two segments, as opposed to classical acoustic models that are designed to discriminate between phonemes that may generate a given segment.

The *Boundary inspector* is a standard feed-forward deep network with two output neurons, which encodes the probability that the central input frame [1] t is a true word boundary. It thus focuses on a single frame, the middle one, and makes a decision about it. It is completed with another model, the *Boundary selector*, which rather considers simultaneously all possible candidate frames in the interval $t \pm 5$, and decides which one is the most likely to be the target word boundary. Because we know that consecutive frames are more correlated than distant frames, we use a recurrent LSTM model to capture correlation between frames. Because no privileged direction is assumed, we use a bi-directional LSTM. The output of this LSTM is then merged with the contextual phonetic information in a feedforward network with 11 outputs: one for each input frame. Both models are finally combined with a deep feedforward network called *Aggregator*, which is trained separately on another corpus.

3.2 Training

Our choice to use a Deep Neural Network (DNN) is motivated by the potential of deep networks to infer complex hierarchical features from data that would have been difficult to design by hand. But this is only possible on large training corpora, while only our small gold corpus is manually annotated with temporal boundaries. We thus have to rely on one of the common deep learning "tricks" for building a large enough training corpus, such as transfer learning, the use of auxiliary tasks or data augmentation. In this

[1] A frame is a time-segment of 10ms length encoded into an acoustic vector of dimension 39 composed of 12 MFCC (Mel-Feature Cepstral Coefficients) plus their derivative and acceleration.

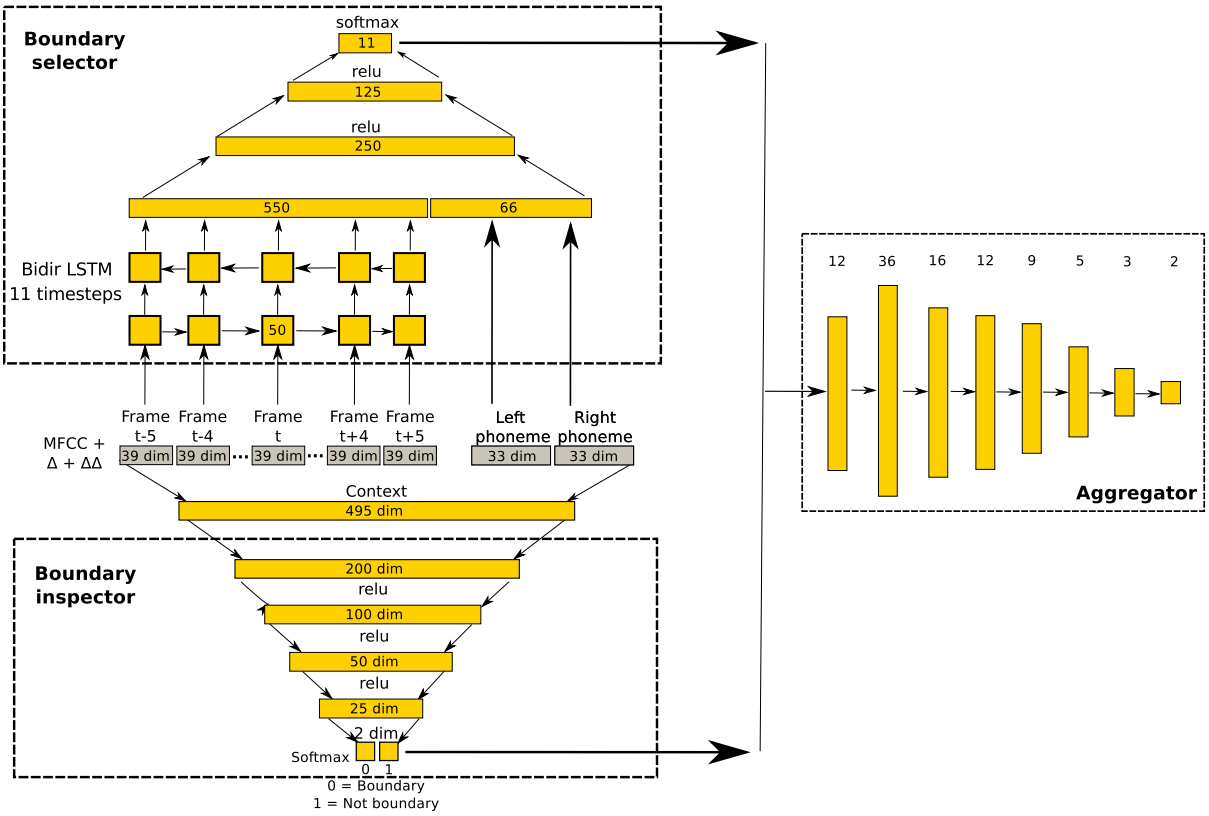

Figure 1: Proposed model. The input vector is composed of 11 frames plus 2 phonemes, shown in the center. Below, the *Boundary inspector* detects whether the middle frame t is a true word boundary or not. Above, the *Boundary selector* is composed of a bi-directional LSTM with 11 timesteps (only 5 are shown) plus 3 feed-forward layers that select the most likely word boundary frame in the input segment. Both models outputs are fed into the *Aggregator* model, on the right, which outputs a confidence probability that frame t is a word boundary.

work, we have decided to combine two state-of-the-art French text-to-speech aligners, ASTALI (Fohr et al., 2015) and JTrans (Cerisara et al., 2009) [2] in order to align part of the open-source ORFEO corpus, composed of 3 million words of French spontaneous speech manually transcribed and available at `http://www.projet-orfeo.fr`.

We then compare both ASTALI and JTrans alignments on this corpus and consider that any word boundary that has the same timestamp in both alignments, within a tolerance of ± 0.02s, is correct[3]. This procedure allows us to automatically build a large training corpus of positive examples which is then completed with 3 times more negative instances obtained by randomly sampling frames that are distant from any ASTALI and JTrans word boundary by at least 0.04s, leading to a training corpus of 377662 examples, which is used to train both the *Boundary inspector* and *selector*.

The same process is used on another set of files from the ORFEO corpus to create a second training corpus of 105406 examples, on which both model output probabilities are computed and used to train the *Aggregator*. During the training of each model, 20% of the training corpus is further reserved to compute a validation loss.

The evolution of the training and validation loss for the three models is shown in Figure 2. These curves suggest that overfitting is not a major issue at this stage. The hyper-parameters of the DNN, including the number and size of the layers, have been set-up empirically with a few trials and errors

[2] JTrans is available on github `https://github.com/synalp/jtrans` and ASTALI is released by its authors

[3] The exact timestamp chosen for this positive temporal limit is the average of the timestamps proposed by JTrans and ASTALI. The tolerance of 0.02s is standard in the phoneme alignment literature (Hosom, 2009).

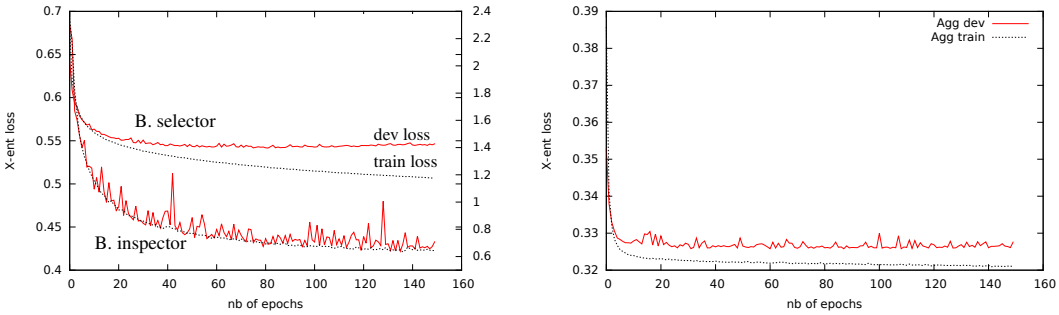

Figure 2: Loss curves during training of the models.

on the training and validation corpus. In particular, we have not used any automatic hyper-parameter tuning strategy. This search of an appropriate model topology has mainly been driven by our motivation to design an architecture that is deep enough to model rich transformations and at the same time that limits the number of parameters to prevent overfitting. Note however that we have only tried a few alternative hyperparameters and thus that the proposed topology is certainly not optimal. The DNN has been implemented with Keras (Chollet, 2015) and trained on these positive and negative instances with the ADAM stochastic gradient descent for 150 epochs.

3.3 Test

The proposed system is evaluated on a gold corpus that is composed of 10988 words extracted from the original ORFEO corpus, and for which 16264 word boundaries, obtained with ASTALI, have been manually corrected. There is no overlapping between this gold corpus and the previous corpora.

At test time, JTrans is run on the test corpus to compute candidate temporal limits of words.

An example of inputs/outputs is shown in Figure 3. Let t be a temporal word boundary[4] given by JTrans, with h_l and h_r respectively the left and right phonemes that are separated by t. For instance, in Figure 3, $t = 186$, h_l =õ and h_r =s.

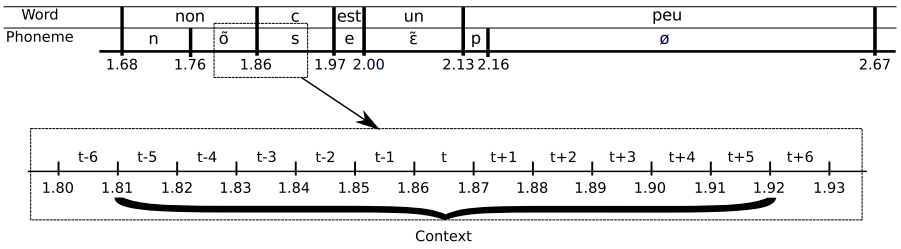

Figure 3: Example of a segmented sentence with its context for the fragment: “no, that’s a bit [...]”

The *Boundary inspector* and *selector* are run on the temporal limits proposed by JTrans and their output probabilities are then passed to the *Aggregator*, which finally returns, for each JTrans temporal limit t, the probability that t is correct or not.

4 Evaluation

4.1 Confidence measure evaluation

Similarly to most other confidence measures in the literature (Yu et al., 2011), we evaluate the proposed confidence measure as a detector of correct vs. erroneous examples. We evaluate next its performances with a Detection Error Tradeoff (DET) curve, which is easier to interpret than the ROC curve (Martin et al., 1997).

[4]Time variables such as t represent an integer number of frames since the start of the audio file. Hence, $t + 2$ is the second frame after the JTrans limit t.

We compare our proposed model first with a baseline confidence measure derived from the acoustic Hidden Markov Models (HMM) used in the text-to-speech alignment process. Let $(w_i)_{1\leq i\leq N}$ be the sequence of words in the transcription. For ease of notation, we assume here without loss of generality that every word is modelled with a single-state Hidden Markov Model (HMM); in fact, every word is actually composed of a sequence of phonemes, and every phoneme is modelled by an HMM with 3 emitting states. However, this hierarchy of models would lead to excessively long equations, and we prefer to simplify the presentation of this baseline.

For a given possible alignment, let the random variable $Q_t = i$ with $1 \leq i \leq N$ represents the index of the word aligned with frame t. By definition, in the context of text-to-speech alignment, a confidence measure for the transition $(Q_t = i, Q_{t+1} = i + 1)$ is given by the posterior probability:

$$P(Q_t = i, Q_{t+1} = i + 1|X, \lambda)$$

where λ represents the parameters of the acoustic models used in JTrans, and $X = (X_t)_{1\leq t\leq T}$ represents all observed acoustic frames.

For our baseline confidence measure, we rely on the acoustic models used in the JTrans system. These acoustic models are Hidden Markov Models, and it is thus well known that the previous posterior can be computed with the forward-backward algorithm:

$$P(Q_t = i, Q_{t+1} = i + 1|X, \lambda) = \frac{\alpha_i(t)a_{i,i+1}\beta_{i+1}(t + 1)b_{i+1}(X_{t+1})}{\sum_{k=1}^{N}\sum_{l=1}^{N}\alpha_k(t)a_{k,l}\beta_l(t + 1)b_l(X_{t+1})}$$

where $b_i(X_t)$ is the observation likelihood of frame X_t in state i. $b_i(X_t)$ is modeled in JTrans by a Gaussian Mixture Model. $a_{i,j} = P(Q_{t+1} = j|Q_t = i)$ is the prior transition probability between words i and j, which is irrelevant in the context of text-to-speech alignment, where we just set $a_{i,i} = \frac{1}{2}$ and $a_{i,i+1} = \frac{1}{2}$. α and β are respectively the matrices of forward and backward probabilities, which can be computed recursively:

$$\alpha_j(t) = \left[\sum_{i=1}^{N}\alpha_i(t - 1)a_{ij}\right] b_j(X_t)$$

$$\beta_i(t) = \sum_{j=1}^{N} a_{ij}b_j(X_{t+1})\beta_j(t + 1)$$

We first evaluate the quality of the DNN as a detector of correct limits, assuming that any JTrans output boundary is a correct limit when its distance to the corresponding gold limit is smaller than 0.02s, as done during training. The corresponding DET curve is shown in Figure 4.

In the DET plot, the closer the curve is to the bottom-left origin, the better it is. We can observe that the proposed confidence measure is a better detector of true boundaries than the acoustic baseline for all possible detection thresholds. The first row in Table 1 also shows the Equal Error Rate (EER), which is the intersection between the $y = x$ diagonal and the DET. With 36% of equal errors, the proposed confidence measure is significantly better than random and it is the first efficient confidence measure for word boundaries based on acoustic information that we are aware of.

While the EER is a good summary of the DET curve, it can only be computed assuming knowledge of the true labels. The next rows in Table 1 thus show standard detection performances at another operating point, the median, which corresponds to the threshold that tags half of the corpus as positive, and half as negative. The proposed model is then compared with a second baseline, called *JTrans/ASTALI agreement*, which tags every JTrans boundary as positive when it lies within the ± 0.02s interval around the corresponding ASTALI limit. This baseline has already been used to automatically annotate the training corpora of the DNN models (see Section 3.2), except that for training, all boundaries tagged as negative are removed, while they are used here to compute the detection metrics in Table 1.

The acoustic baseline confidence measure is not significantly better than random, which confirms for text-to-speech alignment what has already been reported in the literature for speech recognition, i.e., that

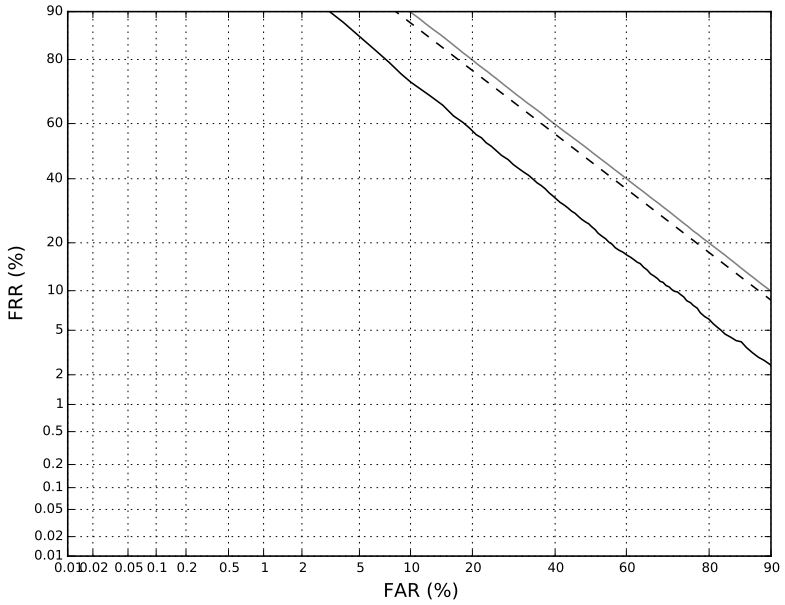

Figure 4: Detection Error Tradeoff curves for detecting word boundaries. The X-axis is the False Accept Rate, while the Y-axis is the False Reject Rate. Three curves are shown, from the worst (top-right corner) to the best (bottom-left corner): random baseline detector (straight grey line), baseline (dash) and our proposed DNN (plain line).

	Acoustic baseline	Proposed model	JTrans / ASTALI agreement
Equal Error Rate (EER)	48%	**36%**	
Precision (median)	43%	52%	60%
Recall (median)	53%	69%	45%
F1 (median)	48%	**60%**	51%

Table 1: Detection performances at fixed operating points of the proposed DNN and two baselines: an acoustic baseline, which computes the posterior of each boundary given JTrans' acoustic models, and a deterministic baseline that tags a boundary as correct when JTrans and ASTALI give close timestamps.

confidence measures based solely on acoustic observation likelihoods usually fail to reliably detect correct words (Willett et al., 1998). This is why state-of-the-art confidence measures for speech recognition mainly exploit other types of features, in particular language-model features (Seigel, 2013). However, language-model information is irrelevant in text-to-speech alignment applications, which makes the task of detecting reliable word boundaries especially challenging. With an F1 of 60%, the performances obtained with our proposed DNN-based confidence measure are thus encouraging, because:

- Our DNN only exploits the same information as speech acoustic models, i.e., acoustic observations and phoneme identities;
- It is trained without manual supervision, only exploiting agreement between two automatic T2SA systems.
- Despite the relatively weak precision of 60% for annotating positive labels in the DNNs' training corpus, the DNN is able to learn relevant acoustic information and provide the first working confidence measure for detecting true word boundaries.

The JTrans/ASTALI agreement baseline has a low recall of 45%. Although this low recall penalizes

its performances as a confidence measure, we can note that the recall is actually not crucial when this JTrans/ASTALI agreement approach is used to automatically annotate training examples for the proposed deep models, because all negative limits are discarded, as explained in Section 3.2. In fact, it may even be preferable to tune this automatic annotation method so that its precision is further increased, at the expense of an even lower recall, so that the positive examples that are kept have a higher likelihood of being correct. However, the JTrans/ASTALI agreement baseline may not easily be tuned in order to increase its precision above 60%. An interesting future work would then be to replace this agreement process with another detector, like the proposed DNN itself, for which the operating point can be tuned, and eventually iteratively retrain the DNN in a self-training fashion on larger corpora, without the need to rely on two different aligners.

4.2 Enhanced aligner evaluation

We propose next a simple post-processing module that enhances the precision of the original T2SA system. This fine-tuning algorithm basically detects suspicious JTrans temporal limits and replaces them by temporal limits with a higher confidence measure in their neighbourhood. It proceeds as follows:

Algorithm 1: Simple fine-tuning of JTrans' output alignment

- For every JTrans output word boundary t:
  - For every distance $d \in 1, 2, 3, 4, 5, \cdots, D$ up to a *maximum distance* D:
    - Compute both DNN output probabilities at distance d from t: $P(t-d)$ and $P(t+d)$
    - Pick the best of both frames $\hat{t} = \arg\max_{t' \in t-d, t+d} P(t')$
    - If the new frame is better than the original JTrans frame $P(\hat{t}) > P(t)$ and better than a minimum confidence threshold $P(\hat{t}) > \delta$, then move the word boundary to $\hat{t}$ and continue with the next boundary t.

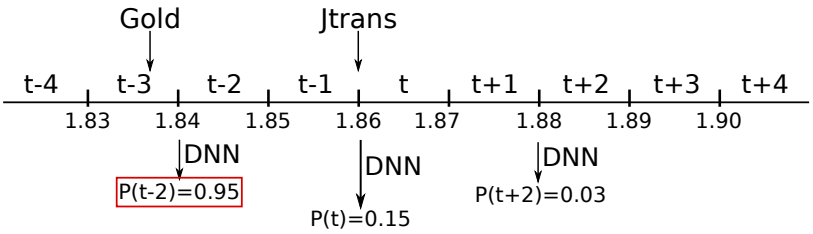

Figure 5: Fine-tuning example. The Jtrans initial limit is 1.86s. The DNN output probability is computed for frames (185,187) first, and then for frames (184,188). The best output is obtained for 1.84s.

Figure 6 plots the original (top horizontal dashed line) and resulting alignment error for various D and δ. Although the global impact of our fine-tuning algorithm is small, it is positive for all D and δ. Because our fine-tuning algorithm just looks for the most confident limits in a neighbourhood of the original JTrans boundary, the iterative application of the confidence measure onto more and more distant frames increases the probability of misclassification. Furthermore, whenever it moves a word boundary, the resulting impact on the previous or following words should be handled, for instance with a Viterbi algorithm. So this experiment is merely a proof of concept that confirms the possibility to post-process a text-to-speech aligner output with the proposed confidence measure; the main focus of this work is rather confidence measure evaluation, which may benefit to many other applications, as discussed in the introduction. These results are thus encouraging to further pursue efforts into investigating weakly supervised deep neural networks for fine-tuning existing text-to-speech aligners.

5 CONCLUSIONS AND FUTURE WORKS

We develop a weakly supervised approach that exploits two existing text-to-speech aligners to automatically annotate a corpus for training a deep neural network-based confidence measure without direct

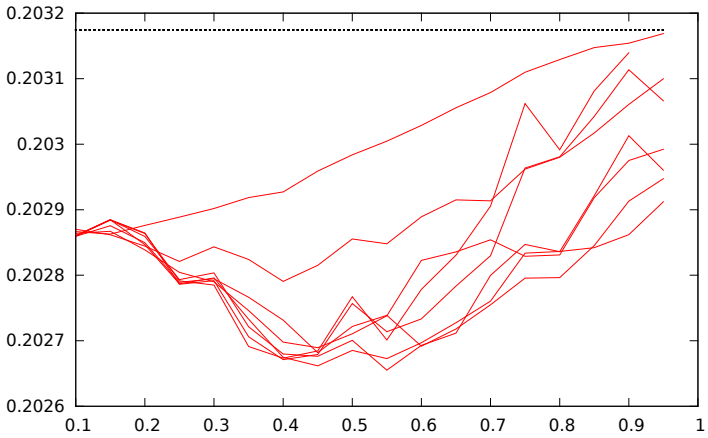

Figure 6: Average alignment error (Y-axis, in seconds) for various maximum distance D (from ± 1 to ± 8 frames), as a function of the minimum confidence threshold (X-axis). The top dashed horizontal line is the original JTrans alignment error.

supervision. We propose two different types of neural networks for this task and combine them within a single model. For now, all three components of the proposed model are trained independently, but we plan to train them jointly in a future work to further improve the resulting model. Experimental results show that the proposed confidence measure outperforms a baseline acoustic confidence measure derived from the original text-to-speech aligner. We further show that it outperforms another baseline, which results from a voting ensemble of both original text-to-speech aligners. This is, to the best of our knowledge, the first good performances ever reported for confidence measure detection of true word boundaries. The performances reached are also interesting because the deep models only exploit acoustic information, which has been shown to be otherwise unsuccessful for confidence estimation in the context of speech recognition, and because these models are trained without manual supervision. These results open the way to further improvements in automatic annotation of unlabelled corpora for text-to-speech alignment, for instance by iteratively re-labelling the training corpus with the proposed model setup in high-precision mode and retraining new confidence models. We further apply the trained confidence measure with a simple corrective algorithm that fine-tunes the output timestamps given by the original text-to-speech aligner. This experiment shows encouraging results for improving text-to-speech alignments thanks to the proposed confidence measure. Possible ways to improve these results include designing a better exploration strategy for fine-tuning the initial alignment, as well as investigating other DNN topologies.

The complete source code as well as links to all datasets is available at `https://github.com/cerisara/speechAlignConfidence`.

Acknowledgments

This work has been partly funded by the ANR ORFEO project. Some experiments presented in this paper have been made possible thanks to the donation of a GPU Titan X card by Nvidia, and were carried out using the Grid'5000 testbed, which is supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

References

B. Bigi. 2013. A phonetization approach for the forced-alignment task. In *Proc. LTC*.

S. Brognaux and T. Drugman. 2016. Hmm-based speech segmentation: Improvements of fully automatic approaches. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1):5 – 15, January.

C. Cerisara, O. Mella, and D. Fohr. 2009. Jtrans, an open-source software for semi-automatic text-to-speech alignment. In *Proc. INTERSPEECH*, Brighton, UK, September.

F. Chollet. 2015. Keras. `https://github.com/fchollet/keras`.

F. de Jong, R. Ordelman, and M. Huijbregts. 2006. Automated speech and audio analysis for semantic access to multimedia. In *Proc. International Conference on Semantic and Digital Media Technologies*, Athens, Greece, December.

D. Fohr, O. Mella, and D. Jouvet. 2015. De l'importance de l'homogénéisation des conventions de transcription pour l'alignement automatique de corpus oraux de parole spontanée. In *8èmes Journées Internationales de Linguistique de Corpus (JLC2015)*, Orléans, France, September.

A. Haubold and J. R. Kender. 2007. Alignment of speech to highly imperfect text transcriptions. In *Proc. IEEE Conf. on Multimedia and Expo*, July.

T. J. Hazen. 2006. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *Proc. Interspeech*, pages 1606–1609.

J.-P. Hosom. 2009. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51(4):352–368, April.

J Keshet, S Shalev-Shwartz, Y Singer, and D. Chazan. 2005. Phoneme alignment based on discriminative learning. In *Proc. Interspeech*, pages 2961–2964.

P. Lanchantin, M. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P. Woodland, and C. Zhang. 2015. The development of the cambridge university alignment systems for the multi-genre broadcast challenge. In *Proc. ASRU*.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. 1997. The det curve in assessment of detection task performance. Technical report, DTIC Document.

B. Milde. 2014. Unsupervised acquisition of acoustic models for speech-to-text alignment. Master's thesis, Univ. Darmstadt, April.

P. J. Moreno, C. Joerg, J.-M. Van Thong, and O. Glickman. 1998. A recursive algorithm for the forced alignment of very long audio segments. In *Proc. ICSLP*, December.

S. Paulo and L. C Oliveira. 2004. Automatic phonetic alignment and its confidence measures. In *Advances in Natural Language Processing*, pages 36–44. Springer.

M. S. Seigel. 2013. *Confidence Estimation for Automatic Speech Recognition Hypotheses*. Ph.D. thesis, Univ. of Cambridge, December.

A. Stan, Y. Mamiya, J. Yamagishi, P. Bell, O. Watts, R.A.J. Clark, and S. King. 2016. Alisa: An automatic lightly supervised speech segmentation and alignment tool. *Computer Speech & Language*, 35:116 – 133.

A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liberman. 2014. Highly accurate phonetic segmentation using boundary correction models and system fusion. In *Proc. ICASSP*, Florence, May. IEEE SPS.

D. Willett, A. Worm, C. Neukirchen, and G. Rigoll. 1998. Confidence measures for hmm-based speech recognition. In *ICSLP*, volume 98, pages 3241–3244. Citeseer.

D. Yu, J. Li, and L. Deng. 2011. Calibration of confidence measures in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2461–2473.

J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang. 2013. Automatic phonetic segmentation using boundary models. In *Proc. Interspeech*, Lyon, August. ISCA - International Speech Communication Association.

Y. Zhao, L. Wang, M. Chu, F. K. Soong, and Z. Cao. 2005. Refining phoneme segmentations using speaker-adaptive context dependent boundary models. In *Proc. Interspeech*, Lisbon, Portugal, September.