Improving Sentence-level Alignment of Speech with Imperfect Transcripts using Utterance Concatenation and VAD

Alexandru Moldovan, Adriana Stan, Mircea Giurgiu
Communications Department
Technical University of Cluj-Napoca
Alexandru.Moldovan@student.utcluj.ro, {Adriana.Stan, Mircea.Giurgiu}@ com.utcluj.ro

***Abstract*—Preparing data for speech processing applications is in general a task which requires expert knowledge and takes up a large amount of time. Therefore, being able to automate as much as possible this process can have a significant impact on the expansion of the number of languages for which spoken interaction with the machines is available.**

In this paper we build upon a previously developed tool, ALISA, which was developed to align speech with imperfect transcripts using only 10 minutes of manually labelled data, in any alphabetic language. Although its error rate is around 0.6% at word-level, we noticed that the sentence-level accuracy is drastically affected by a large number of sentence-initial word deletions. To overcome this problem, we propose two methods: one based on utterance concatenation, and one based on voice activity detection (VAD). The results show that these simple methods can achieve around 10% relative improvement over the baseline results.

***Keywords*—speech and text alignment, VAD, imperfect transcripts, utterance concatenation, ALISA.**

I. Introduction

In speech processing applications, one essential prerequisite is the availability of large amounts of speech data segmented at sentence-level, and having a correct orthographic transcript. However, such data is mostly unavailable for languages outside the mainstream ones (e.g. English, French, Spanish, etc.). Therefore, being able to easily create such data is an important step in developing universal human-computer interaction by means of natural language communication.

There have been a number of approaches to solving this task. These approaches can be broadly classified into two categories: one in which the correct transcript of the speech is known, and for which the goal is to provide the temporal alignment; and one in which the text might contain errors, which also need to be identified and corrected. In the first category the most prominent method is based on either Hidden Markov Model (HMM) acoustic models or dynamic time warping, and it is more commonly referred to as *forced alignment* [1]–[5].

The second category poses additional problems, and generally requires very good acoustic and language models trained on large amounts of supervised data. The main approaches either restrict the language model to match the available transcription [6]–[10], or use acoustic cues to align the speech and text [11].

A major disadvantage of these methods is the fact that they require previously designed expert knowledge, or language specific resources. In our previous work [12], we developed the ALISA tool. ALISA is able to align speech with imperfect transcripts in any alphabetic language starting from just 10 minutes of manually transcribed and annotated speech. The output is on average 70% of the speech data with a sentence error rate (SER) of 7% and a word error rate (WER) of less than 0.5%. These results are achieved by using a highly-restricted language model, called a *skip network* [12], and iterative acoustic model training.

However, when using the results of ALISA to build statistical parametric text-to-speech synthesis systems, a recurring artefact arose from the high number of sentence initial word deletions. This translates into noisy and long duration silence segments at the beginning of the utterance [13]. As a result, in order to improve the results of ALISA even more, these deletions need to be minimised.

In this paper we attempt to reduce the deletions by using two methods: utterance concatenation, and voice activity detection (VAD). The first method performs the alignment of the data by concatenating two neighbouring utterances, and selecting the longest common text subsequence from the recognition output. The use of the skip networks forces the decoding process to perform significantly better within the utterance, as opposed to its ends, and can therefore reduce the number of word deletions obtained for the initial segmentation. The second method uses a GMM-based VAD already available and trained in the initial steps of ALISA. The utterance's starting point as estimated by the VAD is used to select from multiple recognition hypotheses. We show that by using these extremely simple methods, a relative accuracy improvement at the sentence-level of around 10% is achieved.

The paper is organised as follows: Section II presents a brief overview of the ALISA toolkit, and its previously reported results. The proposed methods for SER improvement are described in Section III. Results of these methods are introduced in Section IV, and concluded in Section V.

II. The ALISA Toolkit

ALISA - An Automatic Lightly Supervised Speech Segmentation and Alignment Tool [12] was developed in order to facilitate the production of large amounts of orthographically transcribed speech data, segmented into sentence-like chunks,

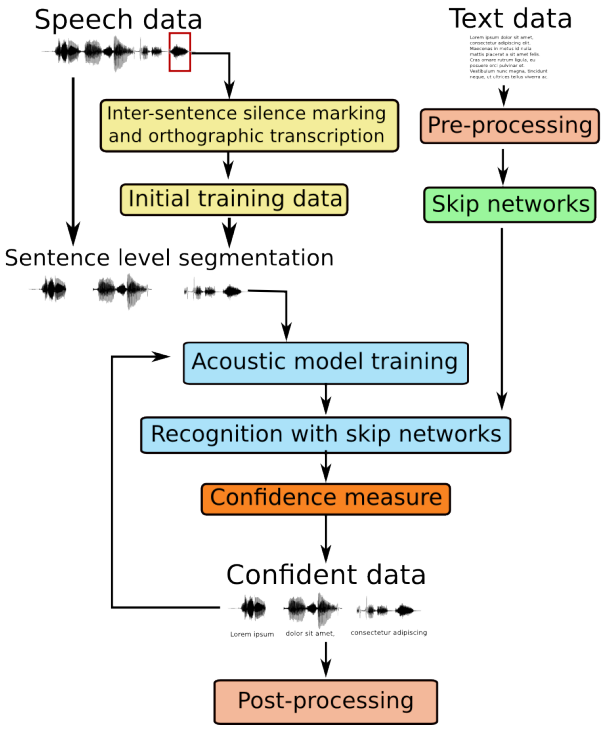

Fig. 1: Flowchart of the ALISA processing steps.

in any language with an alphabetic writing system. The data used by the tool to produce these alignments is mostly readily available, yet not purposely designed for speech applications, such as audiobooks, podcasts, etc. Figure 1 presents an overview chart of all the processing steps involved in ALISA, and the following paragraphs detail the most important blocks.

The alignment starts from only 10 minutes of manually labelled speech. This labelling process refers to the orthographic transcription of the data, and also the manual annotation of the silence segments within it. The silence information is used to train a Gaussian Mixture Model (GMM)-based VAD, which then splits the data into sentence-like segments. Grapheme-level acoustic models are iteratively trained using the initial orthographically transcribed data, and sets of confidently aligned utterances derived from the recognition system. The acoustic model training strategies are incremental and increasingly complex, starting from simple mono-grapheme models, and going to tri-graphemes trained with a discriminative MMI algorithm. In each step, a set of confidently recognised utterances is selected by using a confidence measure based on the recognition acoustic scores.

Due to the fact that the acoustic models are rather poorly trained, and do not benefit from the phonetic information, the recognition is driven by a highly-restricted language model (LM). This LM is built from the available text data, and allows the models to skip only to next words, with a maximum skip of 2 words. This means at most two word deletions. To restrict the decoding even more, a bigram language model trained from the available text data is used to keep just the skip arcs which are a valid bigram within the available text (see Figure 2). Word insertion and substitution is not covered by this type of

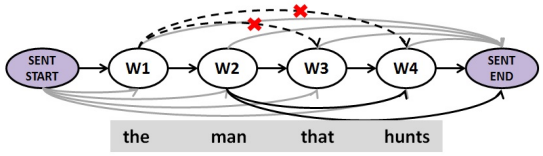

Fig. 2: Skip network with bigram validation.

Recognition output: ~~was~~ a hot and rainy day.
Text-based sentence tokenisation: A hot and rainy day.
(a)

Recognition output: (It) was a hot and rainy day.
Text-based sentence tokenisation: It was a hot and rainy day.
(b)

Fig. 3: Sentence boundary correction method: (a) deletion, (b) insertion.

network.

A post-processing step is also introduced. In it, the recognition output is compared to the text-based sentence-level tokenisation. If there are only slight differences between the two, such as at most two words deleted or inserted, the recognition output is corrected in accordance to the text. We call this process, *sentence boundary correction*, and a sample of such a process is presented in Figure 3.

ALISA's results were objectively and subjectively evaluated in English and French [12]. The error rates are of approximately 7% for sentences, and less than 0.5% for words. The tool was also applied to 14 European languages, and resulted into a multilingual freely available corpus, called Tundra [14].

However, despite the very low WER, the 7% SER poses a serious problem when using ALISA to develop data for text-to-speech synthesis (TTS) systems. Word deletions at the beginning and end of the utterances cause long and noisy silence segments artefacts, as a result of the fact that the silence models are trained on both silence and speech data.

III. Proposed Methods

A. *Utterance concatenation*

As most of the alignment errors made by ALISA are at the beginning and end of an utterance, we wanted to force the decoding process into performing a correct recognition of these segments in particular. One way to achieve this is to concatenate the current utterance with its neighbouring utterances (i.e. the previous and the next one), such that the initial or final speech segments would be in the middle of the newly created utterance. By running the recognition system over these new utterances, the initial speech data's transcript should be contained in both recognition outputs, but will not necessarily have the correct transcript. To extract the transcript of the speech data, a longest common subsequence algorithm is applied over the two outputs. A more explicit description of this process is shown in Figure 4.

There is however a downside to this method. If any of the two neighbouring utterances were not part of the confident

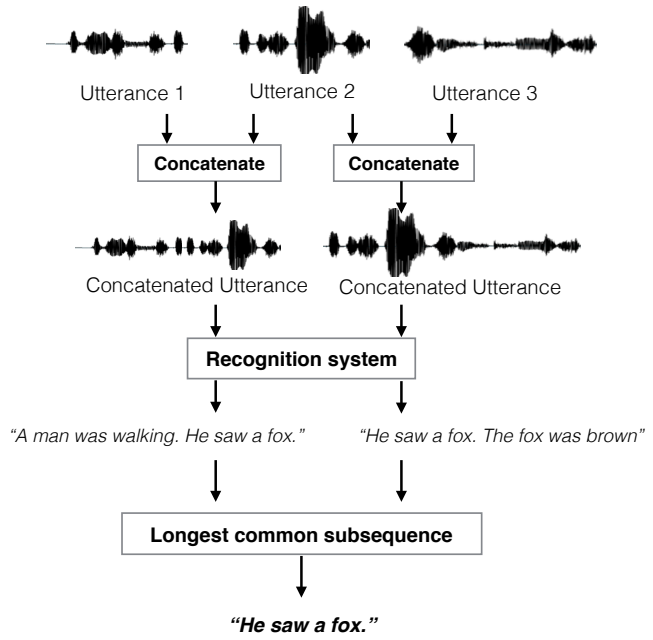

Fig. 4: Utterance concatenation process

set, the entire decoding process could be affected. The result in such a case could be two completely different decoding hypotheses, which would yield an empty transcription, an undesirable effect.

B. VAD-driven Recognition Hypothesis Selection

The ALISA tool also uses HMM-based acoustic models. One of their advantages is the fact that the HMM decoding algorithm can store N-best tokens per state [15], and it can therefore output N-best recognition hypotheses. These hypotheses generally differ only in terms of sentence initial and final mismatches, with very few completely different lattices for the entire utterance. And as such, our word deletion problems could be alleviated by selecting a different hypothesis based on an additional constraint. This constraint could come from a voice activity detector, which would specify, independent of the linguistic content of the utterance, the endpoints of the speech data.

Such a VAD is already available in ALISA, and its default purpose it to segment the input data into sentence-length chunks. The segmentation is based on two GMMs, one for speech, and one for silence. The GMMs are trained on the 10 minutes of manually labelled data. A GMM-based VAD achieves high accuracies due to its data dependent training, as opposed to a general purpose one. As a result, we are able to use the VAD's output as a selection criteria for the recognition hypotheses. And we do this by selecting the hypothesis which has the starting point[1] closer to the one indicated by the VAD labels. The process is simple and efficient, and does not require any additional training material.

[1]Excluding the silence models

TABLE I: Number of deletions and insertions at the beginning, middle, and at the end of an utterance. The figures are computed for the confident utterances set against the GOLD standard transcript.

Position	**No. of deletions [% of total]**	**No. of insertions [% of total]**
Total	342	62
Initial	237 [70%]	7 [12%]
Middle	63 [18%]	26 [42 %]
End	42 [12%]	29 [46%]

IV. Results

A. Data and Acoustic Models

The speech corpus we chose to use in the evaluation of our methods is in the same as in our previously published results [12], and therefore, the improvements are directly comparable. The data is a subset of the 2012 Blizzard Challenge[2] speech corpus, the *A Tramp Abroad* by Mark Twain audiobook. It contains around 15 hours of spoken material, uttered by a professional male speaker, and recorded in a studio at 44.1 kHz, with a 16 bit depth. Gold standard segmentation and transcripts were kindly provided by Toshiba Research Europe Limited, Cambridge Research Laboratory.

From the entire speech corpus, we retained only a subset of around 45% utterances. This subset is the confident data set obtained from the recognition output of the first iteration acoustic models of ALISA. The first iteration models are based on the confident data provided by the models trained on the 10 minutes of manually labelled data.[3] The models are ML, 5-states, left to right, 8 mixture components per state, and no state or mixture tying.

B. Objective Evaluation

We start the evaluation of our methods by analysing the number of word deletion and insertion errors performed by ALISA in the beginning, middle and end of an utterance. Table I shows these errors.[4] It can be noticed that most of the errors in the confident data, approximately 76%,[5] are a result of word deletions. And within this category, 70% of the errors appear in the beginning of the sentence. If we can eliminate these deletions, the WER would be halved.

With this aim in mind, we apply the two proposed methods over the baseline results. Table II presents the SER and WER for the baseline method, as well as for the proposed methods, i.e. utterance concatenation and VAD-driven hypothesis selection. It can be noticed that unfortunately the concatenation lowers the SER by 5%. This was to be expected, especially due to the fact that not all confident utterances are temporally consecutive. Therefore, if any of the neighbouring utterances has a completely erroneous recognition output, this would

[2]http://www.synsig.org/index.php/Blizzard_Challenge_2012.

[3]G1-ML in [12].

[4]There are an additional 75 word substitutions contained in the WER, but because ALISA and this work in particular do not focus on alleviating this type of errors, we will not take them into consideration at this point.

[5]Including the number of word substitutions.

TABLE II: Sentence error rates (SER) and word error rates (WER) for the baseline, utterance concatenation, and VAD-driven hypothesis selection methods.

System	SER[%]	WER [%]
Baseline	11.70	0.6
Concatenation	16.70	0.94
VAD	10.49	0.41

TABLE III: Sentence-initial deletions and insertions for the baseline, utterance concatenation and VAD-driven hypothesis selection methods.

System	Deletions	Insertions
Baseline	237	7
Concatenation	421	6
VAD	110	95

TABLE IV: Number of words deleted at the beginning of a sentence for all three methods.

System	1 word	2 words	>3 words
Baseline	213	19	5
Concatenation	384	33	4
VAD	92	13	5

inevitably affect the transcript of the concatenated utterance as well.

In the case of the VAD-driven hypothesis selection there is some improvement though. The sentence error rate drops by 1.2%, which represents a relative 10%. The WER is also lowered by 0.2%, a 33% relative improvement. If we analyse the results from Table III, we see that the sentence-initial word deletions have been more than halved. However, the downside is that by using this method, a higher number of insertions occur. This is indeed an expected result for the simple hypothesis selection, in which the alignment closer to the one estimated by the VAD algorithm is considered correct.

In Table IV we show a histogram of the number of words deleted at the beginning of an utterance, and it can be noticed that 1, or 2 word deletions prevail. The methods we introduced in this paper are better suited for at most 1 word deletion, and this can be observed by comparing the relative reductions performed by the VAD against the baseline.

V. Conclusions

In this paper we presented two simple methods for improving the sentence-level accuracy for speech and text alignment. The methods are based on utterance concatenation and voice activity detection. Due to the high number of sentence-initial word deletions resulted in the previous version of the tool, the data was not perfectly suited for text-to-speech applications, and resulted into long and noisy duration models.

By imposing additional constraints in the acoustic model decoding process, the number of such deletions was reduced, and therefore the resulting alignments are significantly more accurate. Objective evaluations showed a decrease in the accuracy when using the concatenation method, and a 10% relative improvement when adopting the VAD-driven hypothesis selection. However the increase in word insertions caused by the VAD algorithm requires some future consideration of how this effect could be eliminated.

As future work, we would also like to look into other methods of reducing the word deletions, such as phone-level decoding, or better acoustic model training.

Acknowledgment The research leading to these results has received funding from the Romanian Ministry of Education, under the grant agreement PN-II-PT-PCCA-2013-4 Nº 6/2014 (SWARA).

References

[1] C. Cerisara, O. Mella, and D. Fohr, "JTrans: an open-source software for semi-automatic text-to-speech alignment." in *Proc. of Interspeech*. ISCA, 2009, pp. 1823–1826.

[2] P. J. Moreno, C. F. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Poc. of ICSLP*, 1998.

[3] X. Anguera, N. Perez, A. Urruela, and N. Oliver, "Automatic synchronization of electronic and audio books via TTS alignment and silence filtering," in *Proc. of ICME*, 2011, pp. 1–6.

[4] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *Proc. of Interspeech*, 2007, pp. 2901–2904.

[5] O. Boeffard, L. Charonnat, S. L. Maguer, and D. Lolive, "Towards Fully Automatic Annotation of Audio Books for TTS," in *Proc. of LREC*, Istanbul, Turkey, may 2012.

[6] P. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proc. of ICASSP*, 2009, pp. 4869–4872.

[7] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. of Interspeech*, 2010, pp. 2222–2225.

[8] G. Bordel, M. Peñagarikano, L. J. Rodríguez-Fuentes, and A. Varona, "A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions," in *Proc. of Interspeech*, 2012.

[9] Y. Tao, X. Li, and B. Wu, "A dynamic alignment algorithm for imperfect speech and transcript," *Comput. Sci. Inf. Syst.*, vol. 7, no. 1, pp. 75–84, 2010.

[10] T. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Proc. of Interspeech*, 2006, pp. 1606–1609.

[11] A. Haubold and J. Kender, "Alignment of speech to highly imperfect text transcriptions," in *Multimedia and Expo, 2007 IEEE International Conference on*, July 2007, pp. 224–227.

[12] A. Stan, Y. Mamiya, J. Yamagishi, P. Bell, O. Watts, R. Clark, and S. King, "ALISA: An automatic lightly supervised speech segmentation and alignment tool," *Computer Speech and Language*, vol. 35, pp. 116–133, 2016.

[13] O. Watts, A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, and S. King, "Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from ?found? data: evaluation and analysis," in *Proc. 8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, August 2013.

[14] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, "TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision," in *Proc. Interspeech*, Lyon, France, August 2013, pp. 2331–2335.

[15] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.