Investigating Image Processing Based Aligner for Large Texts

Application for Under-Resourced Languages

Andi BUZO, Horia CUCU, Corneliu BURILEANU
SpeeD Laboratory
University "Politehnica" of Bucharest
Bucharest, Romania
e-mail: {andi.buzo, horia.cucu}@upb.ro, cburileanu@messnet.pub.ro

Abstract—**Speech annotation is a costly and time consuming process because it requires high accuracy. Lightly supervised acoustic modeling solves this problem by making use of approximate transcriptions of speech recordings. In under-resourced languages, the speech recordings are not transcribed entirely and the accuracy of the transcription is poor. In this case, it is necessary an additional segmentation step. We propose a segmentation method that uses image processing techniques in order to spot a text island into a larger one. We also investigate on the effect of several tuning parameters on the method's accuracy.**

Keywords-Lightly Supervised Acoustic Modeling; Text Alignment; Under-resourced Languages; Image Processing.

I. INTRODUCTION

A. Motivation

State-of-the-art Automatic Speech Recognition (ASR) systems are based on large annotated speech data and text corpora in order to train accurate acoustic model and language model. The acoustic model requires hundreds or thousands of hours of annotated speech recordings, while the size of needed text corpora exceeds 1 terra words. The acquisition of such data is a time consuming and costly process. In particular, speech annotation requires higher attention because the speech label must represent the exact speech content. Reformulation of phrases, acronyms, and numbers will degrade the acoustic model accuracy.

Under-resourced languages are characterized by lack of speech and text databases, phonetic dictionaries, lexicons, tools and language expertise. For such languages, web data may represent an important source. However, web data are highly heterogeneous. Speech coding may be different and the sampling frequency may vary from one recording to another. The text content of the speech may not be exactly the same to the one uttered in the recording. Such transcriptions are referred to as loose transcriptions. Non-uniform text formatting should also be addressed through a normalization stage and, for particular languages, even a diacritics restoration stage.

Lightly supervised acoustic model training is largely used in the literature in order to obtain annotated speech databases from loose transcriptions. It is based on an ASR with fair performance that decodes the speech recordings. Then, the ASR output hypothesis is aligned to the loose transcription typically by Dynamic Time Warping (DTW) based techniques. But, the usual scenario in under-resourced languages is the one in which the speech transcription is found in a large text or only a small part of a long recording is transcribed. In this case, the DTW based techniques are not able to align the ASR hypothesis to the text transcription. Hence, a new step must be provided in order to spot the ASR hypothesis within the transcription.

We have proposed such an aligning method in [1]. In this paper, we investigate the effect of the tuning parameters that help increase the alignment accuracy. We also investigate the selection of the confidence scoring that determines the correctly aligned group of words. The rest of the paper is organized as follows: Section II describes the proposed method, Section III presents the experimental setup and the obtained results, while conclusions are given in Section IV.

B. Related Work

In the literature, there are several approaches that overcome the lack of resources barrier. One approach is by adapting the acoustic model of a highly-resourced language with little data from an under-resourced language [2]. Another approach is by building multilingual acoustic and language models as in [3]. Other approaches are focused on obtaining annotated data with little effort. The most common one is lightly supervised acoustic model training.

The first steps in lightly supervised acoustic modeling are made by Moreno et al. in [4]. In this paper, the authors look for confidence text islands called anchors that are supposed to be correctly alignments of ASR hypotheses to loose transcriptions. DTW is used for the alignment, while the confidence text islands are considered to be the ones where the density of the matched words is higher. The anchors are used for segmenting the transcription, so that in a further iteration a specific language model may be used in each segment. The granularity of the transcription is increased at each iteration until convergence is reached. Similar methods are also used in [5] and [6] with small variations in choosing the confidence scoring method. In these methods, strong emphasis is put in biasing the language model towards the transcriptions, which boosts the ASR accuracy.

Another way of matching the ASR hypothesis and the loose transcription is by forced alignment. The main problem with forced alignment is that it assumes that the transcription represents exactly the speech content, which is not the case for loose transcriptions. In [7], Hazen overcomes this drawback by adding a phone loop branch to the decoding network. This extra branch is supposed to absorb the errors that may occur in the transcription, similar to the way out-of-vocabulary words are treated. In [8], the authors solve the same problem by training a *garbage model*.

These approaches are efficient if the size of the ASR hypothesis is comparable to the one of the transcription, otherwise they fail to align them. In [9], the authors attempt to resolve the issue of spotting a text island into a larger text within the context of lightly supervised acoustic modeling. They use the driven decoding algorithm as first iteration where confidence scores are calculated for the recognition hypotheses. These scores are used for language model rescoring in the second iteration.

In the proposed approach, we aim at resolving two main issues with DTW based techniques. The first is the incapability that DTW sometimes has in aligning texts with considerable different sizes. The repetitive words such as common nouns and verbs, prepositions, conjunctions, etc., combined with the poor accuracy of the ASR (especially for under-resourced languages) may create an optimal path that does not necessarily pass through the correct transcription. The second issue is the amount of memory needed for aligning large texts. It may exceed 20 GB if the text has more than 2000 words.

We overcome these issues by proposing a segmentation step whose purpose is to spot the ASR hypothesis (the short text) into the loose transcription (the large text). We achieve this by processing the ASR hypothesis-to-transcription match matrix as a grayscale image. The $m(i,j)$ element of the matrix is 1 if word i of the ASR hypothesis is identical to word j of the loose transcription and 0 otherwise. The matching segment should represent a diagonal line in the image (matrix), thus the method aims at emphasizing this line. This method proved to obtain better results than the DTW based techniques in terms of correctly spotting the target text within a larger one [1].

II. Method Description

A. Lightly Supervised Acoustic Modeling Overview

Lightly supervised acoustic modeling makes use of the loose transcriptions, which approximate the speech content of recordings. Such transcriptions are easier and cheaper to create or to obtain from the Internet. The ASR output hypotheses are obtained by decoding the speech recording using an ASR system with fair accuracy. Intuitively, identical word sequences in the ASR hypothesis and in the transcription suggest that these particular sequences of words are uttered in the respective speech segments. Hence, an alignment of the ASR hypothesis to the transcription is needed in order to localize such regions. Once identified, these segments represent annotated speech that can be used for training a new acoustic model.

In under-resourced languages, lightly supervised acoustic modeling raises supplementary difficulties. First, the accuracy of the ASR is lower, which means that the hypothesis will be more different from the speech content and thus more different from the loose transcription. In this case, the alignment is harder to achieve. Second, the available resources may be highly heterogeneous either from the speech coding or from the text formatting point of view. Third, in many cases only a part of the speech content is transcribed. Under these conditions, the problem consists in *spotting and aligning* a shorter text within a larger similar text. DTW based techniques sometimes fail to achieve this because of the word sequences that are repeated in the texts.

B. The Proposed Method

In this approach, we build a matrix whose elements $m(i,j)$ are 1 if the i^{th} word of the ASR hypothesis is identical to the j^{th} word of the loose transcription and 0 otherwise. Further on, we regard this matrix as a black and white image. If we build such a matrix to align one text to itself, we will obtain an image similar to Fig. 1. The matrix will be *a square one* (NxN, in this example N=60) and its diagonal represents the perfect match of the text's words. The other dots represent the repetitive words in the text. In the general case of aligning two different texts, if the two texts that are going to be aligned are very similar, then the diagonal line will be obvious and it will be easy to spot within the image.

The data collected for under-resourced languages are far from the ideal case. Fig. 2 illustrates the matrix (NxM, in this example N=1700 and M=4200) for the alignment of an ASR hypothesis and a loose transcription. In this image, the diagonal is not obvious because it is "lost" among the high number of dots (representing repetitive words). Fig. 3 shows a zoomed section from Fig. 2 (the white box). It can be observed that the diagonal line is not continuous as it was in Fig.1. This happens because the ASR hypothesis and the loose transcription are not identical. The purpose of our detection method is to make the diagonal line more obvious by using image processing techniques.

The processing steps are the following:

1. ***Repetitive words removal***. The repetitive words whose number exceeds a threshold over the lines, are removed (the whole line is reset to 0). It is possible that these words also occur in the diagonal line, but they do not contribute in emphasizing the diagonal line. On the contrary, they make it less obvious. This condition is expressed mathematically by formula (1):

$$\forall i=\overline{1,N} \quad S_i=\sum_{j=1}^{M} m(i,j) > th_{occ} \quad \Rightarrow \quad \forall j \quad m(i,j)=0 \qquad (1)$$

where $m(i,j)$ represents the element from the row i and column j of the matrix, N is the number of lines, M is the number of columns, S_i is the number of occurrences of the i^{th} word from the ASR hypothesis in the loose transcription, and th_{occ} is the threshold. The result of this processing step is shown in Fig. 4.

2. ***Resolution reduction***. The resolution reduction is made by dividing the matrix into disjunct FxF squares and by calculating the density of each square as shown by equation (2):

$$d(k,l)=\sum_{i=Fk+1}^{F(k+1)} \sum_{j=Fl+1}^{F(l+1)} m(i,j) \qquad (2)$$

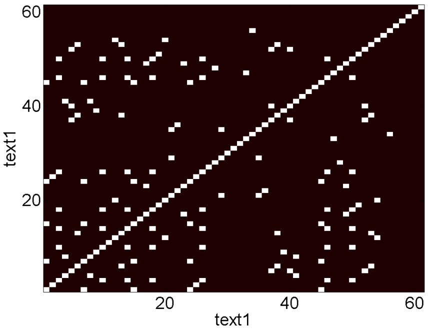

Figure 1. The matrix when aligning one text to itself

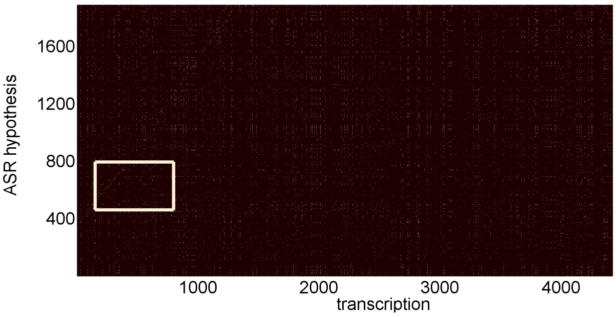

Figure 2. The full matching matrix

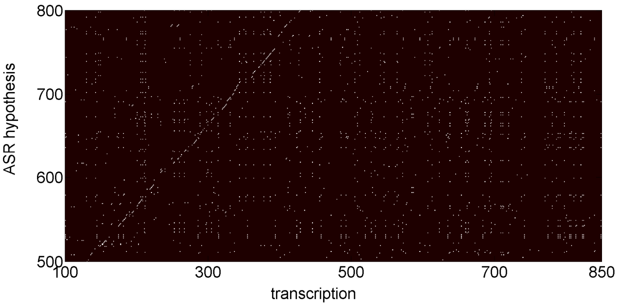

Figure 3. A zoomed region of the matching matrix

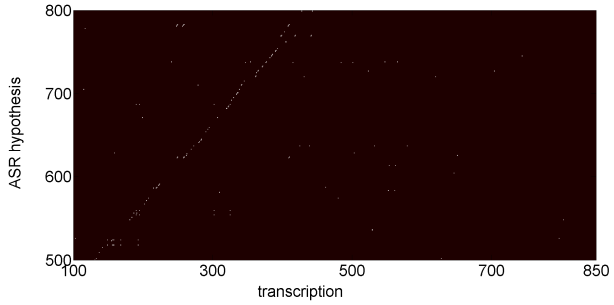

Figure 4. A zoomed region of the matrix after step 1

where $d(k,l)$ is the element of the new matrix d, while F is the factor of resolution reduction. The choice of F has an important impact on the spotting accuracy and its effect is shown by the experimental results (see Section III).

This processing step has a double role. First, it emphasizes the diagonal line because it is expected that the rest of the left dots are spread in an irregular way over the image (see Fig. 5). Second, thanks to the lower resolution, the computational load will be lower in the following steps.

3. ***Wiener filtering***. A 2D Wiener filtering is a low pass filter that reduces the additive noise power. The removal of the repetitive words at step 1 reduces the correlation among dots making their distribution more noisy. Without step 1 the Wiener filtering will not be successful. The filtering is performed pixel-wise according to equation (3) [10]:

$$\hat{d}(k_1,k_2) = \mu + \frac{\sigma^2 - \upsilon^2}{\sigma^2}(d(k_1,k_2) - \mu) \qquad (3)$$

where:

$$\mu = \frac{1}{K^2}\sum_{l_1-K/2}^{l_1+K/2}\sum_{l_2-K/2}^{l_2+K/2} d(l_1,l_2) \qquad (4)$$

$$\sigma^2 = \frac{1}{K^2}\sum_{l_1-K/2}^{l_1+K/2}\sum_{l_2-K/2}^{l_2+K/2} d(l_1,l_2) - \mu^2 \qquad (5)$$

The mean (μ) and the variance (σ^2) are calculated within a square KxK matrix in the neighborhood of each pixel. υ^2 is the average of all variances. K determines how much the dots or the contours are faded, thus it must be chosen carefully.

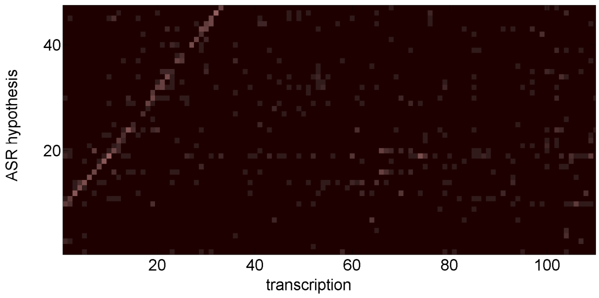

Figure 5. The full matrix after resolution reduction

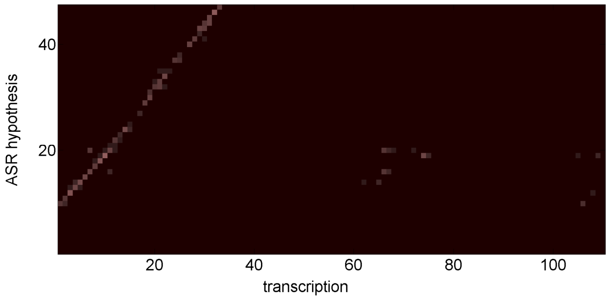

Figure 6. The matrix after Wiener filtering (step 3)

The experimental results in Section III show how the method efficiency depends on *K*. The result of this processing step is shown in Fig. 6.

4. ***Rough alignment***. At this step, we spot the ASR hypothesis within the loose transcription using intercorrelation. We do this by calculating the maximum values over the lines (max_l) and over the columns (max_c). It is expected that vectors max_l and max_c are highly correlated. The index for which the cross-correlation of the two vectors obtains the maximum value, provides the offset of the alignment for the ASR hypothesis and loose transcription.

C. *Fine Alignment and Confidence Scoring*

After the rough alignment, the word alignment is made by using a standard DTW technique that is used also for the calculation of Word Error Rate (WER). After the words are aligned a confidence score is needed in order to determine whether the matched words are indeed uttered in the speech content. Because the performance of the ASR system is not high we cannot count on the output likelihood. Instead, we make the assumption that longer matched words and longer matched word sequences are more probable to represent the content of the speech recording. The acceptance rule for the matched words is the following: words longer than *m* characters and sequences longer than *n* words are considered correct matches. Thresholds *m* and *n* strongly depend on the performance of the ASR system and they are determined empirically (see Section III).

III. EXPERIMENTAL RESULTS

A. *Experimental Setup*

Romanian language is considered to be an under-resourced language because there are no annotated speech resources available, neither free nor paid ones. Hence, the database used for lightly supervised acoustic modeling is collected from Romanian news websites. However, the collected speech recordings are not transcribed entirely. In other cases, the transcription of a short speech recording is found within a larger text. In both cases, segmentation is needed before word alignment. The accuracy of the transcriptions is not high. There is a WER of 30% compared to the exact speech content. The errors consist in several tags, special number formats, re-phrasing of speech content, misspellings, etc. The database used for experiments contains 100 hours of broadcast news.

The experiments are made with an ASR system in the Romanian language [11], [12]. The acoustic model is trained with 70h of read speech from 80 speakers. It is based on 3-state Hidden Markov Models, while the number of senones is 4000. Each state's output is modeled by a Gaussian Mixture Model with 64 densities. The vector of speech features has 39 elements (13 Mel Frequency Cepstral Coeficients + Δ + $\Delta\Delta$). The language model is 3-gram based and it is trained with 170 million of words. The ASR system obtains 20% WER for read clean speech and 59% WER for broadcast news.

B. *Refinement of Method's Parameters*

The purpose of these experiments is to determine the optimal values for the tuning parameters described in Section II. These parameters are the threshold of removal (th_{occ}) for the occurrences of repetitive words, the resolution reduction factor (*F*) and the size of the neighborhood in Wiener filtering (*K*). By varying these parameters a huge number of combinations can be obtained. Because the segmentation algorithm is time consuming (20h of run for 100h of speech), we have decided to maintain two parameters fixed while varying the third. The metric used for the accuracy of the segmentation method is the number of hours of the extracted annotated speech.

Table I shows how the method's efficiency varies with th_{occ}. The results show that there is an optimal value for th_{occ}. Indeed, th_{occ} cannot be too high, otherwise the remaining dots from step 1 will not have a noisy distribution, but rather a well defined pattern. In this case, the Wiener filtering is not efficient. On the other hand, th_{occ} cannot be too low either because words from the diagonal line will also be removed making it more subtle.

The effect of parameter *F* over the method's performance is shown in Table II. The results confirm that higher values of *F* make the diagonal line more obvious and easier to spot leading to a larger amount of extracted speech. This happens because the density of the dots along the diagonal line is higher than the density of the other areas that contain sporadic dots. However, *F* cannot be too high because for some transcriptions the number of words is smaller and this imposes a limit on the resolution reduction.

The effect of parameter *K* over the method's performance is shown in Table III. A higher *K* leads to a greater fading of the dots and a greater smoothening of the contours (the diagonal line). This contributes in making the diagonal line more obvious among the dots. For K>5 no further improvement is obtained. For experiments, we will choose K=5 as the optimal value because higher values for *K* involves more computations.

C. *Annotated Speech Extraction*

For the rough alignment of the ASR hypotheses and the loose transcriptions we have used the optimal values determined in Section III.B: th_{occ}=3, F=40, K=5. Once the texts are aligned, word alignment is performed based on a DTW based technique. In order to evaluate the performance of the system for several values of *m* and *n* described in Section II.C, we have verified in each case 10 minutes of annotated speech resulted from word alignment process. The results are shown in Table IV. The number in each cell indicates the percentage of speech segments that are erroneously annotated. Errors in speech annotation lead to inaccurate acoustic models. Hence, *m* and *n* are chosen so that no errors occur in the automatic annotation process (m=8 and n=4).

TABLE I. THE EFFECT OF TH_{OCC} IN SPEECH EXTRACTION

th_{occ}	***F*=20, *K*=3**
	No. extracted speech hours
2	4,2
3	6,2
4	6,1
5	5,9
6	5,7
7	5,2

TABLE II. THE EFFECT OF *F* IN SPEECH EXTRACTION

F	**th_{occ} =3, *K*=3**
	No. extracted speech hours
10	5,5
20	5,7
30	5,9
40	6
50	5,9

TABLE III. THE EFFECT OF *K* IN SPEECH EXTRACTION

K	**th_{occ} =3, *F*=40**
	No. extracted speech hours
2	5,9
3	6
4	6,1
5	6,2
6	6,2
7	6,2

TABLE IV. THE ERROR PERCENTAGE FOR DIFFERENT EXTRACTION RULES

n* \ *m	***5***	***6***	***7***	***8***
2	3.0%	2.9%	2,1%	1.1%
3	2.0%	1.7%	1.5%	1.4%
4	0.9%	0.7%	0.3%	0.0%
5	0.5%	0.4%	0.3%	0.0%

IV. CONCLUSIONS

Lightly supervised acoustic modeling is a cost-effective method of obtaining annotated speech data. However, traditional DTW-based techniques sometimes fail to align large texts. The proposed method overcomes this issue by introducing a segmentation step based on image processing. Its purpose is to emphasize a diagonal line, which represents the correctly matched words, among other dots that represent the repetitive words.

The method is more efficient if the dots have a noisy distribution, thus the repetitive words whose occurrence number exceeds a threshold are removed. The experimental results also proved that there exist optimal values for other two tuning parameters: the resolution reduction factor and the size of the Wiener filtering neighborhood. By these refinements, the method outperforms the traditional DTW based aligning methods.

REFERENCES

[1] A. Buzo, H. Cucu, and C. Burileanu, "Text spotting in large speech databases for under-resourced languages," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing. Canada, *submitted paper*, 2013.

[2] V-B. Le and L. Besacier. "First steps in fast acoustic modeling for a new target language. Application to Vietnamese," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing. USA, pp. 821 - 824, 2005.

[3] D. Yu, L. Deng, P. Liu, J. Wu, Y. Gong, and A. Acero, "Cross-lingual speech recognition under runtime resource constraints," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, Taiwan, pp. 4193-4196, 2009.

[4] P. J. Moreno, C. Joerg, J.-M. Van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in Proceedings of ICSLP, USA, pp. 2711-2714, 1998.

[5] L. Lamel, J. Gauvain, and G. Adda, "Investigating Lightly Supervised Acoustic Model Training," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, USA, pp. 189-192 vol. 1, 2001.

[6] H. Y. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, Canada, pp. 737-740 vol.1, 2004.

[7] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," Proceedings of Interspeech, USA, pp. 1606-1609, 2006.

[8] M. H. Davel, C. Van Heerden, N. Kleynhans, and E. Barnard, "Efficient harvesting of Internet audio for resource-scarce ASR," Proceedings of Interspeech, Italy, pp. 3154-3157, 2011.

[9] B. Lecouteux, G. Linares, J.F. Bonastre, and P. Nocera, "Imperfect transcript driven speech recognition," Proceedings of Interspeech, USA, pp. 1626-1629, 2006.

[10] Lim, Jae S., "Two-Dimensional Signal and Image Processing", Englewood Cliffs, NJ, Prentice Hall, 1990, pp. 536-540.

[11] H. Cucu, L. Besacier, A. Buzo, and C. Burileanu, "Enhancing Automatic Speech Recognition for Romanian by Using Machine Translated and Web-based Text Corpora," Proceedings of SPECOM 2011, Rusia pp. 81-88, 2011.

[12] H. Cucu, L. Besacier, C. Burileanu, and A. Buzo, "Investigating the Role of Machine Translated Text in ASR Domain Adaptation: Unsupervised and Semi-supervised Methods," in the Proceedings of the 2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), USA, pp. 260-265, 2011, ISBN: 978-1-4673-0366-8.