Text Spotting In Large Speech Databases For Under-Resourced Languages

Andi Buzo, Horia Cucu, Corneliu Burileanu
Speech and Dialogue (SpeeD) Research Laboratory
University "Politehnica" of Bucharest
Bucharest, Romania
{andi.buzo, horia.cucu, corneliu.burileanu}@upb.ro

***Abstract*—Lightly supervised acoustic modeling in under-resourced languages raises new issues due to the poor accuracy of Automatic Speech Recognition (ASR) systems for such languages and the quality of the speech transcriptions that may be found. In these conditions, the common alignment techniques are not always capable of aligning the ASR output and the approximate transcription. We propose two aligning methods that overcome these issues. In the first approach we apply an image processing algorithm on the matching matrix of the two texts to be aligned, while the second alignment approach is based on segmental DTW. The approaches outperform the current Dynamic Time Warping technique (DTW) by extracting in average 29% and 27% respectively more speech data than the currently used DTW.**

Keywords— lightly supervised acoustic modeling; text alignment; under-resourced languages

I. INTRODUCTION

State-of-the-art Automatic Speech Recognition (ASR) systems for high-resourced languages use hundreds or even thousands of hours of annotated speech recordings for training the acoustic model (AM) and corpora with billions of words to train the language model (LM). The collection of such data is expensive and requires a lot of time. Under-resourced languages are characterized by lack of text and annotated speech data, phonetic dictionaries, tools and language expertise. Consequently, there is no straight-forward way of developing robust AMs and LMs for these languages. One way to overcome this barrier is by deriving new algorithms capable of adapting AMs and LMs from a highly-resourced language to an under-resourced language [1], [2].

Another way is to provide new methods of automatic data acquisition. Such a method is the lightly supervised AM training, which is based on a low performance ASR system used as bootstrap and approximate transcriptions of speech recordings. Approximate or loose transcriptions are easier to find (e.g. transcribed interviews, movie captions, etc.) and cheaper to obtain. The key issues in lightly supervised AM training are: (1) segmentation of data into homogeneous parts, (2) alignment of the approximate transcription to the ASR output and (3) confidence scoring of the aligned text in order to filter the relevant data.

In this paper, we present a study whose aim is to enrich the annotated speech database for the Romanian language, which is an under-resourced language from the acoustic modeling point of view. We propose new unsupervised methods for the alignment of loose transcriptions to the ASR output. The evaluation is made on a Romanian database of approximate transcriptions acquired semi-automatically from the Internet. The speech recordings contain broadcasted news, talk shows, interviews, etc. Eventually, the aligned transcriptions are used for training the AM.

The remainder of the paper is organized in four sections. Section 2 presents the most relevant related works, Section 3 describes the proposed methods and Section 4 evaluates them. Finally in Section 5 we draw the conclusions of this work.

II. RELATED WORK

In the last decade, lightly supervised acoustic modeling has been an interesting topic to many research groups because of its capability to provide annotated data in a cheaper way. A common way is by applying Dynamic Time Warping (DTW) for aligning the loose transcriptions and the ASR output and look for perfectly matched words [3, 4, 5, 6]. In [3], the authors provide a confidence score for the matched words and then use the sequences of words with the highest score as anchors. These anchors are used for segmenting the texts and for each obtained segment a specific LM is build in order to increase the ASR accuracy. The process is repeated by increasing the granularity of the segmentation in each iteration. This method is not suitable for ASR systems with poor recognition accuracy. Moreover, if the text to be aligned is very large then the repetitive expressions could make the anchors ambiguous and consequently unusable.

Forced alignment is another way for aligning the approximate transcriptions and the ASR output [7, 8]. A pure forced alignment is risky because it assumes a strong confidence scoring that distinguishes the correct matched words from the rest of the words that are simply forced to align. In [7], Hazen uses a pseudo forced alignment, i.e. he assumes that errors may occur in the transcriptions or during recognition and consequently allows the insertion or the substitution with other words by the means of a phone loop out-of-vocabulary (OOV) word filler. In [8], the authors use phone level forced alignment and a garbage model for OOV

words. The standard Word Error Rate (WER) procedure is used as confidence scoring by biasing towards the matched words.

A big issue in the alignment process is the execution time, because all these methods run several iterations of ASR. Lecouteux et al. propose a method that combines the search with the driven decoding algorithm [9]. First, they find matching anchors by using DTW and then they use the matching scores for linguistic rescoring. In [10], the same authors reduce the execution time by merging the two steps.

Due to the intrinsic characteristics of the DTW algorithm, these methods are efficient only if two conditions are fulfilled: (1) the size of the transcription segments that are going to be processed must fit the computer capabilities. This involves an accurate segmentation method. (2) In order to reduce the execution time, the search graph or the paths are pruned. This optimization fails if there is high temporal mismatch between the transcriptions and the ASR output (e.g. large missing parts). For under-resourced languages none of these conditions is fulfilled. The usual scenario is a short transcribed part from a long audio or a short audio whose transcription is found in a larger text.

We go beyond the state-of-the-art barrier by proposing two aligning algorithms for spotting a text island in a larger text that the DTW used in [3-6] fails to spot. The first is a novel method that uses a matrix representation of the match between the transcription and the ASR output. The matrix whose values can either be 1 for match and 0 for mismatch, is processed as an image. Besides a diagonal line which is expected to represent the correct alignment, the other dots in the image, which represent the incorrectly matched words, are filtered by a 2D Wiener filter [11]. Eventually, after the diagonal is localized, the text is segmented properly. The second method is based on segmental DTW that we have previously used in Spoken Term Detection [12]. The DTW is applied on a sliding window whose length is proportional to the shorter text (usually the ASR output). WER is used as the detection score in order to spot the correct text segment.

In order to improve the alignment and the recognition accuracy we also use LM bias towards the transcriptions as suggested in [4-6]. We use a similar confidence score as the one suggested in [3], i.e. sequences of n or more consecutive aligned words are considered to be correctly recognized. In addition, the matching isolated words that have more than m phonemes are also considered correct matches (*n* and *m* are determined empirically).

III. The Description of the Methods

A. Overview of the Lightly Supervised Acoustic Training

Our lightly supervised AM training includes the following summarized steps:

1. Data acquisition. Several news websites contain an audio/video version of the news besides the text version. Such data are collected automatically facilitated by the presence in HTML of the hyperlink to the audio/video content. In other cases, the speech and the text data are obtained from different sources.

2. Data normalization. Text data are normalized using some previously developed procedures [13]. The diacritics are restored, numbers and dates are converted into text, several tags are removed, etc. Audio/video data are grouped by their coding and then converted to a common coding: 16KHz, PCM, 16bits/sample.

3. ASR. The speech recordings are decoded using an ASR system that we have previously developed in [13, 14]. The LM is biased toward the transcriptions in order to increase the recognition accuracy on this particular data.

4. Data segmentation and alignment. Usually the transcription is larger than the output of the ASR, thus it is necessary to spot the ASR hypotheses in the large transcript. For this we use the methods described in Section 3.2 and 3.3. Then, the standard DTW used for the calculation of WER aligns the two texts. Sequences of at least n aligned words and single aligned words of more than m phonemes are considered correctly matched. In the experiments, we have used n=4 and m=8, which were determined empirically.

5. AM training. The correctly matched data are used for training a new AM.

Steps 3-5 are repeated until the number of correctly matched words does not increase significantly from one iteration to the next. These steps are fully automated.

B. Image-Processing-based Alignment (IPA)

In this approach, we represent the match between the transcription and the ASR hypothesis by an *N*x*M* matrix where *N* is equal to the number of words in the recognition hypothesis and *M* is equal to the number of words in the transcription. The matrix values are 1 for the correct matched words and 0 for the incorrect ones. This matrix is further processed as a black & white image.

Matches (in the transcription and ASR hypothesis) for frequent, repetitive words (such as “the”, “from”, “is”, etc.) will be represented as repetitive single white dots (having the same abscise or the same ordinate), while matches for repetitive words sequences (such as “point of view”, “he is going to”, “it is not”, etc.) will be represented as short parallel diagonals. A match for a long word sequence (in the transcription and ASR hypothesis) will be represented as a white diagonal line in this image.

Because of the poor accuracy of the ASR system and the approximate transcriptions (which do not reflect the exact words uttered in the recording), the matching diagonal line is not obvious (see Figure 1). Figure 2 is a zoom in Figure 1 (where the diagonal line is less obvious). Such images are further processed to emphasize the diagonal line and thus spot the ASR hypothesis in the transcription.

The processing stages are the following:

1. Repetitive words filtering. If a word from the ASR hypothesis is found more than a threshold number in the transcription, the respective line from the matrix is set to 0 (see expression 1).

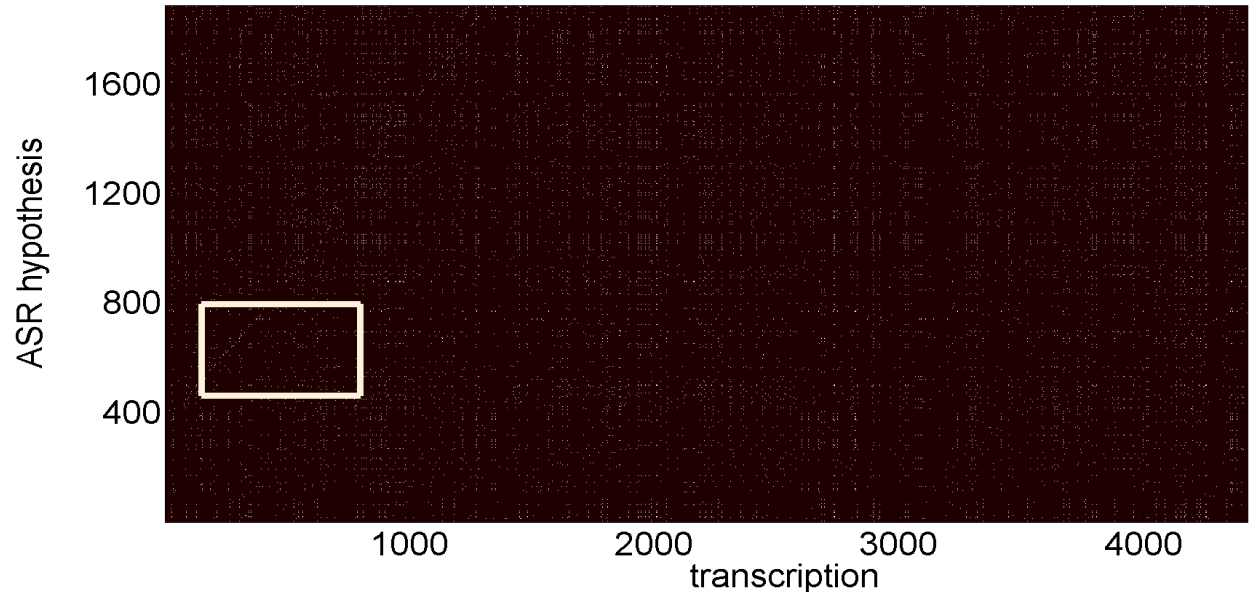

Fig. 1. The full matching matrix

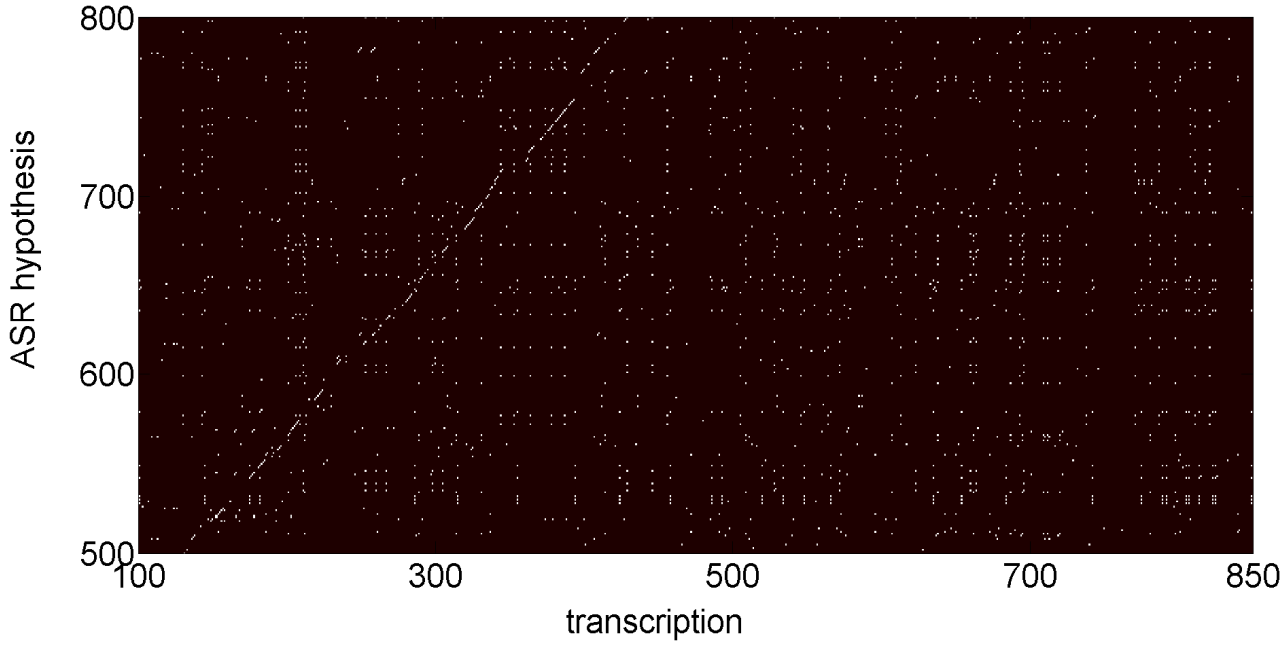

Fig. 2. A zoomed region of the matching matrix

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

where m is the NxM matrix, S_i is the sum over line i and th_{occ} is the threshold of occurrences for a word. The same rule is applied for columns too. The result is illustrated in Figure 2.

2. Resolution reduction. The resolution is reduced with a factor F by calculating the densities on disjunct squares of FxF as shown in equation (2):

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

where m is the matrix modified at step 1, and d is the matrix in low resolution (Figure 4).

3. Wiener filtering. A Wiener filter is applied on the matrix d obtained at step 2. This is a low-pass filter that reduces the additive constant power noise [11] (Figure 5).

4. Rough alignment. After filtering, it is expected that the greatest values of matrix d are found on the diagonal line. For this reason, the vector of maximums over lines (*maxL*) is highly correlated to the vector of maximums over columns (*maxC)*. Thus, for the localization of the ASR hypothesis within the transcription, we align *maxL* and *maxC* by using the correlation function. The maximum value of the correlation shows the offset between ASR hypothesis and the transcription. The transcription is segmented to match the length of the ASR hypothesis.

5. DTW. The segmented transcription and the ASR hypothesis are aligned by using a standard DTW as in the calculation of WER.

6. Matched words selection. Each sequence of at least n correctly matched words and the words with more than m phonemes are selected along with their corresponding speech signal for training a new AM (see Section 3.1.).

In order to have an efficient Wiener filtering, the dots (besides the diagonal line) must have a noisy character. Instead, they have a rather static pattern, i.e. for a common word all the dots are on the same line or column (Figure 2). After the removal of such words at step 1, the remaining dots are "randomly" distributed in the image (Figure 3). Step 2 has a double function: (1) it makes the diagonal line denser, hence immune to Wiener filtering (note: the application of Wiener filtering after step 1 obtained poor results because the dots from diagonal line were also vanished) and (2) it reduces the resolution which means that fewer points are processed in the next steps.

In our experiments, we have used th_{occ}=3, F=40 and K=5. These values were determined empirically and proved to be effective in emphasizing the diagonal line.

C. Segmental DTW (S-DTW) Alignment

This method is common in pattern recognition and we have previously used it in Spoken Term Detection [12]. We define a sliding window with the size proportional to the length of the ASR hypothesis. The overlap between two consecutive windows is chosen 66% in order to assure the coverage of any relative position of the ASR hypothesis against the transcription. For each window the DTW is applied between the ASR hypothesis and the respective transcription segment and WER is calculated. It is expected that the segment with the lowest WER belongs to the speech recording used in ASR. Eventually, the aligned words are selected by the rule described in Section 3.A.

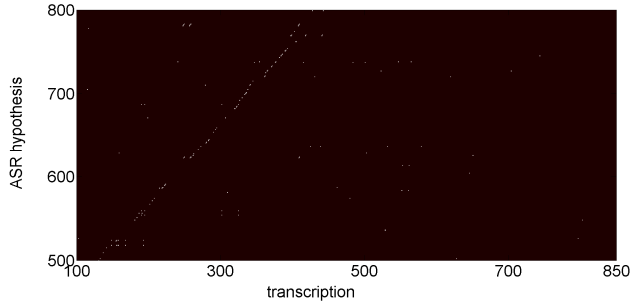

Fig. 3. A zoomed region of the matrix after step 2

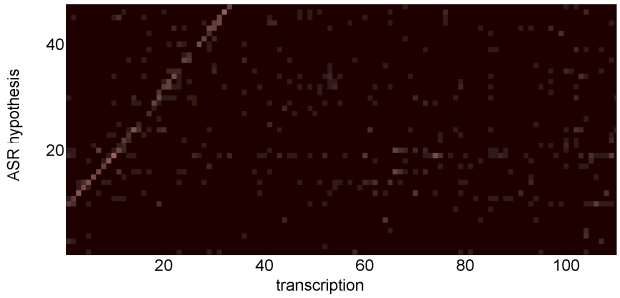

Fig. 4. The full matrix after resolution reduction

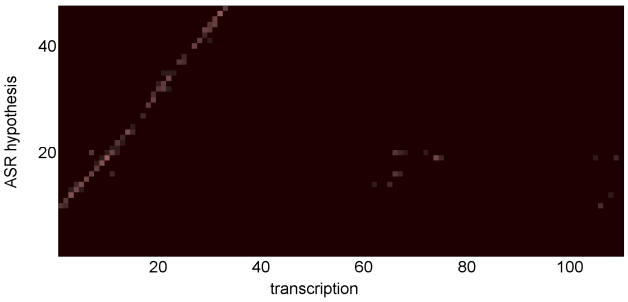

Fig. 5. The matrix after Wiener filtering (step 4)

IV. Experimental Results and Discussion

A. Experimental Setup

The collected approximate transcriptions and speech recordings are very heterogeneous. Most of them are partially transcribed interviews, broadcast news, talk shows, etc., taken from different websites that use different audio coding. The transcription is poor, especially from the ASR point of view; rephrasing is very often used, diacritics are sometimes missing and out-of-vocabulary words are frequent because the text may belong to any domain. In average, a transcription has 30% WER compared to the exact speech content. Often, the relevant transcriptions are only a small part of a larger text. Among the 400h of Romanian speech we collected so far only 100h

contain relevant speech. These 100h out of 400h were chosen for the comparison experiments done in this paper. 3h of perfectly annotated speech are used for measuring the accuracy of the improved AM.

The ASR used for recognition was previously developed by our research group [13, 14]. Hidden Markov Models are used for acoustic modeling with 3 states per phoneme and a total of 4000 senones. The state's output is modeled using Gaussian Mixture Models with 64 components. The AM is trained with 70h of read speech. 39 coefficients (13 MFCC+Δ+ΔΔ) are used as speech features. The LM is n-gram-based and it is trained with a corpus of 170 million words. The ASR performance varies from 20% WER on read, clean speech to 59% WER on broadcast news.

B. Performance evaluation

The proposed methods: Image-Processing-Alignment (IPA) and Segmental DTW (S-DTW) are compared with DTW from the alignment efficiency and processing time points of view. The metric used for the alignment efficiency is the amount of correct aligned speech, measured in hours. The evaluation is made on the test database of 100h. The performance of DTW is highly affected by the heterogeneity level of the transcription and by the length difference of ASR hypothesis and transcription. The greater the heterogeneity level and the length difference, the poorer the DTW's capability to align the texts will be. For this reason, the speech recordings from the test database are artificially segmented into clips of 4, 8, 16, 32 and 64 minutes in order to have different levels of heterogeneity and different hypothesis-to-transcription length ratios (HTLR). For example, if an 8 minutes ASR output is aligned to a 1h transcription, the HTLR is about 1:8. In our tests, for each level of segmentation, the ASR hypotheses are aligned to transcriptions corresponding to 1h of speech.

The results presented in Table I show that S-DTW and IPA outperform DTW especially for low HTLR. In this case, DTW fails to align some clips due to the poor accuracy of ASR, incorrect transcriptions and repetitive words in large transcriptions (1h of speech contains approximately 2000 words), thus the DTW optimal path may pass through the repetitive words instead of the correct ones. In our database, the average HTLR is 1:4. For this HTLR, S-DTW and IPA obtain 27% and 29% respectively much more aligned speech than DTW.

The computation time for S-DTW increases with smaller HTLR because the number of iterations for the sliding window is greater. IPA introduces a fixed overhead for the segmentation part (while for word alignment, DTW is used) which does not depend on HTLR (the overhead is 21h of processing for 100h of speech). However, for lower HTLR, the DTW processing time is shorter in IPA because the transcriptions are trunked, thus there are less data to align. The computation time is very important in this process because of the huge amount of data that has to be processed. However, the segmentation by IPA or S-DTW occurs only at the first iteration of alignment, hence the overhead introduced is reasonable.

TABLE I. THE PERFORMANCE OF THE METHODS

Clip duration [mins] / HTLR	Extracted speech [h]			Processing time [h]		
	DTW	S-DTW	IPA	DTW	S-DTW	IPA
4 / 1:16	3.8	5.3	5.3	8.1	22.8	21.5
8 / 1:8	4.1	5.8	5.9	8.1	21.3	22.0
16 / 1:4	**4.4**	**6.0**	**6.2**	**8.1**	**18.2**	**23.0**
32 / 1:2	5.4	6.0	6.1	8.1	12.2	25.0
64 / 1:1	6.1	n/a	n/a	8.1	n/a	n/a

TABLE II. THE ALIGNMENT EFFICIENCY OF IPA

	Extracted speech [h]	Overall efficiency [%]	Aligned efficiency [%]
Baseline	15,3	3,8	15,3
Iteration 1	19,6	4,9	19,6
Iteration 2	30,2	7,5	30,2
Iteration 3	32,0	8,0	32,0

TABLE III. THE ACCURACY IMPROVEMENT OF THE ASR SYSTEM

	WER [%]
Baseline	59.09
Baseline with biased LM	56.69
Baseline with interpolated LM	55.49
After the 1st iteration with interpolated LM	50.26
After the 2nd iteration with interpolated LM	47.19
After the 3rd iteration with interpolated LM	46.27

Next, we have used IPA for alignment and the extraction of the annotated speech data that are further used in training a new AM as explained in Section 3.1. The alignment is run over the whole database of 400h. The amount of annotated speech obtained at each iteration is presented in Table 2. Note that after the first iterations the amount of extracted speech increases because the baseline AM is trained with read speech and the introduction of spontaneous speech in the training data boosts the ASR accuracy. However, after 3 iterations the gain is low because no significant new data are introduced in the training database.

Table II also shows that after all the iterations, only 8% of the collected speech data can used for acoustic modeling. However, from the total of 400h, not all the recordings contain relevant speech. After filtering the irrelevant data, we measured 100h of relevant recordings. The last column of Table II shows the efficiency relative to the these 100 hours of relevant recordings. Hence, the speech extraction efficiency on this data is 32%.

The overall ASR improvement is presented in Table III. The evaluation is made on 3 hours of perfectly annotated spontaneous speech from broadcast news. This database represents unseen data for the ASR as they are not used in the AM improvement process. The use of the biased LM reduces the WER significantly. A higher reduction of WER is obtained if an interpolated LM is used. The interpolation is obtained from the general LM and the news LM with bias on news LM. As the AM is trained with more and more speech data, the WER is further reduced. However, the most significant reduction is obtained after the first AM improvement iteration. The overall WER reduction is 22% relative.

The speech extraction efficiency and the WER reduction show that the alignment methods can obtain annotated data with little effort, but for a better ASR accuracy, more speech data must be collected.

V. CONCLUSIONS

Lightly supervised AM training presents new issues when it is applied to under-resourced languages. These issues are the poor accuracy of the initial ASR in such languages and the highly heterogeneous speech data that may be found. These lead to highly dissimilar ASR hypotheses and transcriptions that need to be aligned. Under these conditions, the DTW-based methods used in the literature are not always capable of aligning the two texts.

The two proposed approaches overcome these issues because they focus on spotting the relevant text first and then performing the alignment. The experimental results showed that by IPA and S-DTW we obtain in average 29% and 27% respectively more annotated speech data than DTW. Both methods introduce a computation overhead, but because these methods are run only once, it becomes insignificant.

From 400h of data we were able to extract 32h of annotated speech which is used in the AM training. The WER of the ASR trained with the newly acquired data is reduced by 22% relative.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] L. Nguyen and B. Xiang, "Light supervision in acoustic model training," ICASSP: Proceedings of the Acoustics, Speech, and Signal Processing, Canada, pp. 185-188 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] B. Lecouteux, G. Linares, F. Beaugendre and P. Nocera, "Text island spotting in large speech databases," Proceedings of Interspeech, Belgium, pp. 1318-1321, 2007.

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

[12] A. Buzo, H. Cucu, M. Safta, B. Ionescu and C. Burileanu, "ARF @ MediaEval 2012: A Romanian ASR-based Approach to Spoken Term Detection," Working Notes Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-927/mediaeval2012_submission_18.pdf.

[13] 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.

[14] 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.