ARF @ MediaEval 2012: A Romanian ASR-based Approach to Spoken Term Detection

Andi Buzo[1]
andi.buzo@upb.ro

Horia Cucu[1]
horia.cucu@upb.ro

Mihai Safta[1]
cmsafta@upb.com

Bogdan Ionescu[2]
bionescu@imag.pub.ro

Corneliu Burileanu[1]
cburileanu@messnet.pub.ro

ABSTRACT

In this paper, we attempt to resolve the Spoken Term Detection problem for under-resourced languages within the Automatic Speech Recognition (ASR) paradigm. The proposed methods are validated with unseen dataset in African languages.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Speech Recognition and Synthesis – *keyword spotting, spoken term detection*.

Keywords

Spoken Term Detection, Automatic Speech Recognition.

1. INTRODUCTION AND APPROACH

We approach the Spoken Web Search Task @ MediaEval 2012 [1] starting from an ASR system for the Romanian language. The task involves searching for audio content within audio content using an audio query. The ASR is adapted to minimize the Phone Error Rate (PhER) for the Lwazi development corpus (http://www.meraka.org.za/lwazi/pd.php). The ASR system is completed with a search block which processes various types of the ASR outputs (the best string hypothesis, lattice and confusion sets).

1.1 Proposed Romanian ASR

For this task, we use the ASR system for the Romanian language that we have previously developed and described in [2]. The acoustic model is build using 64 hours of speech from different speakers. Its best performance is 18% Word Error Rate (WER) with a language model trained with 170 million words. In order to reduce the mismatch between the Lwazi (test) the Romanian (training) database, we have filtered the Romanian speech recordings to 8 KHz. The available Lwazi development data set has a small lexicon, hence the usage of speech recognition at word level leads to a high number of out-of-vocabulary words and consequently to a high WER. For this reason, we have chosen phone recognition and tuned (relative beam width, word insertion probability, language weight, etc.) the Romanian ASR to minimize the Phone Error Rate (PhER).

1.2 ASR adaptation

The Lwazi data set consists of audio content recorded over telephone channels in four of the 11 South African languages. Audio content consists of a combination of read and elicited speech. In total, there are approximately 1,500 items, plus about 100 spoken search queries. The Romanian language and the South African languages do not have the same phones. For this reason, we performed a mapping between the two phone sets. 77 African phones are mapped to 28 of the Romanian phones (the Romanian language has 34 phones). Each African phone is mapped to a Romanian phone by applying the rules listed below, in this specific order:

a. If its IPA classification is identical to the IPA classification of a Romanian phone, the phones are mapped directly.
b. Else, the closest Romanian phone is found by using the full IPA chart (http://www.langsci.ucl.ac.uk/ipa/fullchart.html).
c. All recordings from the Lwazi developing set are transcribed using the Romanian ASR for phones. A confusion matrix for the phones of both languages is built based on the ASR output and the real transcriptions. If none of the first two rules is applicable, then the mapping is made according to the confusion matrix.

A maximum a priori (MAP) adaptation is performed to the acoustic models for the Romanian phones by using the Lwazi developing set. The unigram language model used for ASR is constructed using the counts of the African phone occurrences found on the Lwazi website. At this point, all the queries and the contents are transcribed by using the adapted ASR and they are passed to the search block for Spoken Term Detection (STD).

1.3 Searching techniques

We designed and experimented with several searching techniques, which can be divided into two categories:

a. ***Techniques based on character comparison.*** If the ASR accuracy would be 100% then the STD is reduced to a simple character string search of a query within a textual content. As the experimental results show (48% PhER), we are far from the ideal case, hence we have to find within a content a string which is *similar* to the query. The search of the *exact* query string has poor STD results: 94% Miss Proability (MP) and 0.4% False Alarm Probability (FAP). Moreover, it does not offer the possibility to find a compromise between MP and FAP.

The DTW String Search (DTWSS) uses the Dynamic Time Warping to align a string (a query) within a content. The search is not performed on the entire content, but only on a part of it by the means of a sliding window proportional to the length of the query. The term is considered detected if the DTW scores above a threshold. This method is refined by introducing a *penalization* for the short queries and the spread of the DTW match. The formula for the score s is given by equation (1):

$$s = (1 - PhER)(1 + \alpha \frac{L_Q - L_{Qm}}{L_{QM} - L_{Qm}})(1 + \beta \frac{L_W - L_S}{L_Q}) \qquad (1)$$

where L_Q is the length of the query, L_{QM}=17 and L_{Qm}=4 are the maximum and the minimum query lengths found in the development data set, L_W is the length of the sliding window, L_S

[1] SpeeD, University Politehnica of Bucharest, Romania.

[2] LAPI, University Politehnica of Bucharest, Romania

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

is the length of the matched term in the content, while α and β are the tuning parameters.

The *Sausage technique* (denoted ST) is based on the confusion networks built on the lattice output of the ASR [3]. Thus, the content is a chain (*sausage*) of confusion sets (CS). Each CS has one or more phones with the respective transition probability. The alignment of the query with the sausage is made as in DTWSS by weighting the score according to the spread of the alignment and the length of the query.

b. ***Techniques based on acoustics.*** These techniques base their decisions on the log likelihood probability that a speech recording obtains in ASR. The lattice obtained from the recognition of a query is used to build a finite state grammar for the ASR. The contents are then passed to an ASR process that uses the *Lattice Grammar* (LG). The grammar has a loop edge so that a long content can pass through the lattice multiple times. It is expected that if the content contains the query term they will obtain a higher score while passing through the LG.

2. EXPERIMENTAL RESULTS

2.1 ASR accuracy

We started from the Romanian ASR which had PhER of 36.8% (tested on Romanian data). After tuning the beam width related parameters we succeeded in reducing PhER to 31.4%. The tuning of language related parameters (language weight and word insertion penalty) brought a further reduction of PhER to 25.3%. We have split the African developing data set into training (90%) and testing (10%). At this point, the recognition of the African testing data set with the Romanian acoustic model obtained a PhER of 61.2%. We performed a MAP adaptation of the Romanian acoustic model by using the African training data set and obtained a PhER of 50.3% when using the word labels and a PhER of 48.1% when using the phone labels. It is the latter acoustic model that we have used for STD.

2.2 STD results and official runs

The results obtained in the official runs for all methods on the evaluation data set are shown in Figure 1. The primary metric used for comparison is Actual Term Weighted Value (ATWV). It is obvious that the DTWSS method is by far the most efficient. The effect of weighting the score according to the query length and the spread of the alignment match is given by the results presented in Table 1. These results are obtained for a sliding window length equal to 1.5 times the query length. The shorter the query, the greater are the chances that different, but similar words obtain higher scores. This is why better results are obtained by giving α a greater value. Similarly, the greater the spread of the DTW match, the lower the probability that it is the searched term. However, there is an optimal value for both α and β. The parameters for the 3 DTWSS official runs are chosen based on these values.

Table 1. DTWSS results

ATWV	α=0	α=0.1	α=0.2	α=0.4	α=0.6	α=0.8	α=1
β=0	0.21	0.21	0.22	0.22	0.23	0.22	0.22
β=0.2	0.29	0.29	0.31	0.30	0.32	0.28	0.25
β=0.4	0.31	0.33	0.33	0.33	0.33	0.34	0.30
β=0.6	0.31	0.32	0.32	0.32	0.33	0.31	0.31
β=0.8	0.28	0.31	0.30	0.32	0.30	0.28	0.26

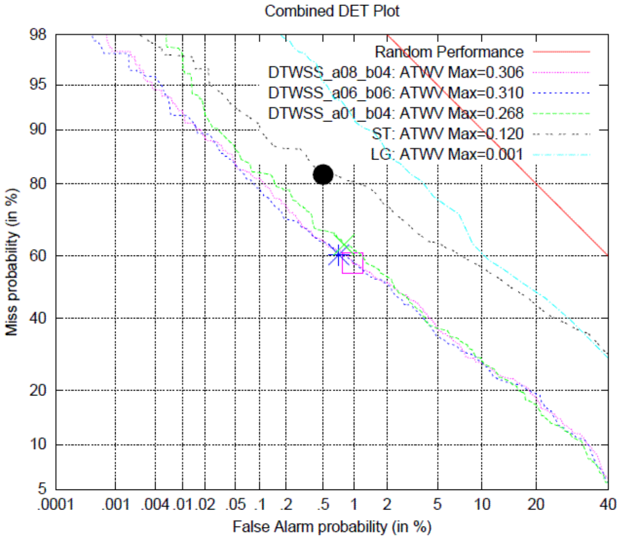

Figure 1. The official run results for the eval data set

The ST and the LG methods are based on the recognition hypotheses alternatives. We expected that these methods would decrease significantly the MP, but due to the poor recognition accuracy (PhER of 48%) the hypotheses alternatives are very different from the true one, hence leading to high FAP.

The official runs results for all combinations of testing data sets (dev/eval queries, dev/eval contents) are shown in Table 2. All the methods suffer performance degradation when moving from training data to unseen data. However, the degradation is not drastic. The *evalQ-devC* combination obtains better results than the *devQ-devC* one, because in the latter case the amount of unseen data (the contents) is higher.

Table 2. Official run results

ATWV	*evalQ-evalC*	*evalQ -devC*	*devQ-evalC*	*devQ-devC*
DTWSS (α=0.8 β=0.4)	0.31	0.47	0.33	0.49
DTWSS (α=0.6 β=0.6)	0.31	0.48	0.33	0.47
DTWSS (α=0.1 β=0.4)	0.27	0.44	0.32	0.47
ST	0.12	0.22	0.17	0.25
LG	0	0.02	0	-

3. CONCLUSION

We have based STD on a Romanian ASR adapted for the African languages. We tested various searching methods and DTWSS obtained the best results. The results are improved if the decision score is weighted according to the length of the query and the spread of the alignment match.

4. REFERENCES

[1] F. Metze, E. Barnard, M. Davel, C. van Heerden, X. Anguera, G. Gravier, N. Rajput, ” The Spoken Web Search Task”, MediaEval Workshop, Pisa, Italy, 4-5 October 2012.

[2] H. Cucu, L. Besacier, C. Burileanu, A. Buzo, ”Enhancing Automatic Speech Recognition for Romanian by Using Machine Translated and Web-based Text Corpora,” SPECOM 2011, pp. 81-88, Kazan, Russia, 2011.

[3] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: Word error minimization and other applications of confusion networks,” Computer Speech and Language, vol. 14, no. 4, pp. 373–400, October 2000.