# SpeeD @ MediaEval 2013: A Phone Recognition Approach to Spoken Term Detection

Andi Buzo[1] andi.buzo@upb.ro
Horia Cucu[1] horia.cucu@upb.ro
Iris Molnar[1] iris.molnar@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 multiple languages.

## 1. INTRODUCTION AND APPROACH

We approach the Spoken Web Search (SWS) Task @ MediaEval 2013 [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 used for converting the speech signal into a string of phonemes. This indexing process is supposed to run offline. During runtime a searching block finds in the content database the matches for the query string that have a similitude score above a given threshold.

### 1.1 The 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 SWS database and the Romanian (training) database, we have filtered the Romanian speech recordings to 8 KHz. The SWS database does not have transcriptions of any kind. For this reason, it is not feasible to use word recognition. Phone recognition is used instead. Phones are common across many languages. The Romanian language has only 26 phonemes, whereas the languages from SWS database count at least triple that number. The choice of having less classes than phonemes is not so hazardous as many multiple language ASR systems use phoneme clustering by grouping acoustically similar phonemes into the same class.

The trained Hidden Markov Models divide speech feature space into phoneme classes and phonemes that do not belong to the Romanian language will be classified into one of these classes. If all instances of an out-of-Romanian-language phoneme are recognized as a given Romanian phoneme, than the problem is solved, because both queries and contents will have the same symbol. However, this is not always the case and the experimental results confirm this. It seems that many phonemes occupy a region in the speech feature space that overlap only partially with the Romanian phonemes subspaces. For the Romanian database the ASR system yields a Phone Error Rate (PhER) of 36.8%.

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).

### 1.2 Searching algorithm

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, 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: 99% Miss Proability (MP) and 0.1% 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):

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

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, $L_S$ is the length of the matched term in the content, while $\alpha$ and $\beta$ are the tuning parameters.

The penalizations in formula (1) are motivated by the assumption that for two queries of different length that match their respective contents by the same PhER, the match of the longer query is more probable to be the right one. Similarly the more compact DTW matches are assumed to be more probable than the longer ones.

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


*MediaEval 2013 Workshop*, October 18-19, 2012, Barcelona, Spain

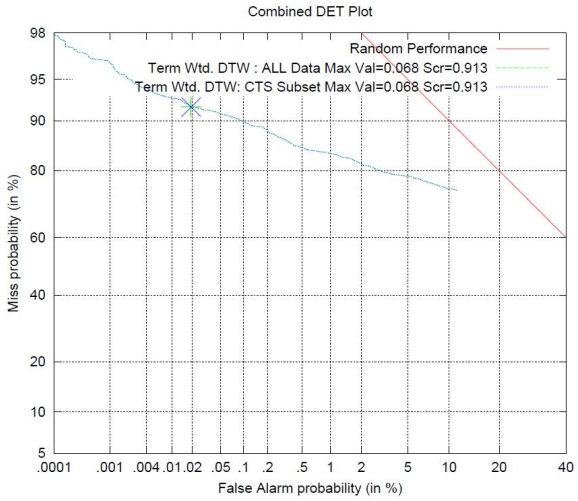


**Figure 1. The primary run results for the eval data**

# 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%.

## 2.2 STD results and official runs

The results obtained in the official run for the primary method on the evaluation data set is shown in Figure 1. The primary metric used for comparison is Maximum Term Weighted Value (MTWV). For other values of $\alpha$ and $\beta$ similar curves are obtained. Depending on the application a trade-off can be made between false alarms and miss rates.

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 which are obtained with the development data. 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 $\alpha$ 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 $\alpha$ and $\beta$. The optimal values for $\alpha$ and $\beta$ are explained by the fact that by increasing $\alpha$ above a certain value the shorter queries are given higher scores just for being short, even though they might have a great PhER value. The same rationale takes place for the spread of the DTW match for which the $\beta$ factor is responsible. The parameters for the 3 DTWSS official runs are chosen based on these values.

**Table 1. DTWSS results**

| ATWV | $\alpha$=0 | $\alpha$=0.1 | $\alpha$=0.2 | $\alpha$=0.4 | $\alpha$=0.6 | $\alpha$=0.8 | $\alpha$=1 | $\alpha$=1.2 |
|---|---|---|---|---|---|---|---|---|
| $\beta$=0 | 0.050 | 0.056 | 0.059 | 0.064 | 0.066 | 0.065 | 0.065 | 0.064 |
| $\beta$=0.2 | 0.052 | 0.058 | 0.062 | 0.064 | 0.066 | 0.067 | 0.066 | 0.066 |
| $\beta$=0.4 | 0.052 | 0.058 | 0.062 | 0.065 | **0.068** | 0.067 | 0.066 | 0.066 |
| $\beta$=0.6 | 0.052 | 0.060 | 0.062 | 0.065 | 0.067 | **0.068** | 0.066 | **0.066** |
| $\beta$=0.8 | 0.052 | 0.060 | 0.062 | 0.065 | 0.067 | 0.067 | 0.066 | 0.066 |

The official runs results for all combinations of testing data sets (development data, evaluation data) 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. This is explained by the fact that development data are used only for tuning the $\alpha$ and $\beta$ parameters but not for adapting the ASR system. Overall, the results are poor which means that the trained Romanian phonemes do not divide optimally the speech feature space. Phonemes from other languages lie in regions between two or more classes and different instances of the same phoneme are classified differently, thus increasing the uncertainty in the decision process. The system can be significantly improved phonemes from other languages can be trained and introduced in the phone recognizer.

Given the simplicity of the searching component the real time factor is very low ($6*10^{-5}$ $s^{-1}$). Memory used during the process is 7.1 GB.

**Table 2. Official run results**

| MTWV | *Dev data* | *Eval data* |
|---|---|---|
| DTWSS ($\alpha$=0.6 $\beta$=0.4) | 0.068 | 0.058 |
| DTWSS ($\alpha$=0.8 $\beta$=0.6) | 0.068 | 0.059 |
| DTWSS ($\alpha$=1.2 $\beta$=0.6) | 0.066 | 0.053 |

# 3. CONCLUSIONS

We have approached STD with a two step process. A Romanian ASR is used as a phone recognizer for indexing the database, while a DTW based algorithm is used for searching a given query in the content database. 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] X. Anguera, F. Metze, A. Buzo, I. Szoke and L.J. Rodriguez-Fuentes, "The Spoken Web Search Task", MediaEval 2013 Workshop, 18-19 October 2013, Barcelona, Spain.

[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.