SpeeD @ MediaEval 2014: Spoken Term Detection with Robust Multilingual Phone Recognition

Andi Buzo, Horia Cucu, Corneliu Burileanu
SpeeD Research Laboratory, University Politehnica of Bucharest
{andi.buzo, horia.cucu, corneliu.burileanu}@upb.ro

ABSTRACT

In this paper, we attempt to resolve the Spoken Term Detection (STD) problem for under-resourced languages by phone recognition with a multilingual acoustic model of three languages (Albanian, English and Romanian). The Power Normalized Cepstral Coefficients (PNCC) features are used for improved robustness to noise.

1. INTRODUCTION AND APPROACH

We approach the Query by Example Search on Speech Task (QUESST) @ MediaEval 2014 [1] by using a multilingual acoustic model (AM) trained with three languages (Albanian, English and Romanian). The task involves searching for audio content within audio content using an audio query. The approach consists in two stages: (1) the indexing, i.e. the phone recognition of the content data and (2) the searching, i.e. finding a similar string of phones in the indexed content that matches the one of the query by using a DTW based searching algorithm.

1.1 The acoustic model

In our approach, we want to compare the effect of using multilingual AM against the monolingual AM. In order to achieve this we have built five acoustic models described in Table 1. The AM training and the phoneme recognition are made by using Hidden Markov Models (HMMs).

Table 1. Training data

ID	Language	No. phonemes	Training data [h]
AM1	Romanian	34	8.7
AM2	Albanian	36	4.1
AM3	English	75	3.9
AM4	Multilingual separate phones	145	16.7
AM5	Multilingual common phones	98	16.7
AM6	Romanian MediaEval 2013	34	64

We have built an AM for each language, (AM1 - AM3). AM1 is trained with 8.7 hours of read speech. We had more available training data for Romanian (in the MediaEval 2013 evaluation campaign we used 64 hours [2]), but this year we chose to train with less Romanian data in order to have a balanced training data set among different languages. AM2 is trained with 4.1 hours of Albanian read speech and broadcast news. AM3 is trained with 3.9 hours of native English read speech from the standard TIMIT database [3]. All these three languages are part of the languages used in MediaEval 2014 evaluation campaign [1] (except for English which is non-native). Hence, using more training data would go beyond the context of the competition which aims at low-resourced languages. AM4 is trained with all the data from the three languages. Phonemes from different languages, however, are trained separately. This led to a big number of phonemes (145). AM5 was trained with the same data as AM4, but in contrast phonemes that are common in different languages were trained together, thus reducing the number of phonemes to 98, which is still high. The identification of the common phonemes was made based on International Phonetic Alphabet (IPA) classification [4]. It is interesting to notice that Romanian and Albanian had in common more than 80% of their phonemes. As for English, it has in common many consonants with the other two languages, but very different vowels. AM6 is the one used by SpeeD team in MediaEval 2013 and it is tracked here for comparison [2].
Two speech features types are used in this work: the common Mel Frequency Cepstral Coefficients (MFCC) and the Power Normalized Cepstral Coefficients (PNCC).

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 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}=18 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 α and β are the tuning parameters. In this work, α and β are set to 0.6.
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 phone error rate (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. This algorithm is suitable for queries of type 1 and 2, because the DTW handles inherently the small variations from the query, but it is not suitable for queries of type 3 where words order may be inverted.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

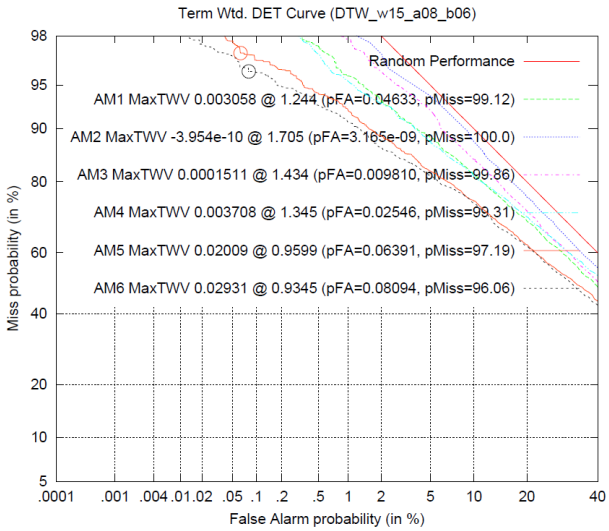

Figure 1. The results for the development data set

2. EXPERIMENTAL RESULTS

2.1 STD results

The results obtained with different acoustic models on the development data set are shown in Figure 1. The comparison is made by using the Maximum Term-Weighted Value (MTWV) and the Detection Error Tradeoff (DET) curves. The speech features used are the PNCCs. By comparing the acoustic models trained with a single language, the Romanian AM outperformed the other two. This is most probably because the Romanian AM is trained on more data (8.7h vs. ~4h). AM4 performed slightly better than the monolingual acoustic models. On one hand, it is trained with multiple languages which would increase the phoneme recognition accuracy, on the other hand the number of phonemes for this acoustic model is significantly increased which increases the uncertainty during recognition. AM5 improves this latter aspect by not training separately common phonemes among different languages and the results show an improvement in performance. However the best results are obtained with AM6. Even though it is trained with only one language (Romanian), it is trained with a big amount of data (64h) and the set of phonemes is relatively small (34). This means that for larger phonemes set larger data are needed for training. Regarding the STD task, it seems that by training with multiple languages the performance increases but more data are needed in order to consolidate the acoustic models.

Table 2. PNCC vs. MFCC performance comparison

ID	PNCC		MFCC	
	ACnxe	MinCnxe	ACnxe	MinCnxe
AM1	1.032	0.986	1.032	0.986
AM2	1.055	0.997	1.055	0.997
AM3	1.03	0.994	1.03	0.994
AM4	1.015	0.972	1.016	0.971
AM5	1.016	0.969	1.016	0.969

The results obtained on the development database with different speech features (PNCC and MFCC) are shown in Table II. The metric used is the *normalized cross entropy cost* (Cnxe). The results show almost no difference between the two types of features. The same conclusion is drawn even when comparing by TWV metric. In general speech recognition, PNCCs obtain better accuracy in noise conditions, but, most probably, the noise in the MediaEval 2014 database is not significant. Therefore, the use of PNCC did not bring any improvement.

2.2 Official runs results

The results obtained by the official runs on the evaluation database are shown in Table 3 and the metrics used are the actual and the minimum Cnxe. Because no tuning is made based on the development data set, the results on the evaluation data set are quite similar and the same conclusions can be drawn. Table 3 shows also the results per query type. It can be noticed that better results are obtained by query type 2. In contrast to query type 1, these queries are longer, which may have affected the results. Query type 3 has obtained a slightly worse performance, most probably because of the reordering of the words in such queries.

Table 3. Official runs

	Overall A/Min Cnxe	Type 1 A/Min Cnxe	Type 2 A/Min Cnxe	Type 3 A/Min Cnxe
AM1	1.032/0.990	1.035/0.990	1.027/0.982	1.039/0.992
AM2	1.053/0.997	1.057/0.999	1.046/0.994	1.052/0.995
AM3	1.027/0.990	1.029/0.991	1.024/0.983	1.032/0.994
AM4	1.017/0.977	1.019/0.976	1.012/0.973	1.018/0.974
AM5	1.017/0.972	1.019/0.972	1.016/0970	1.017/0.963

The results are obtained on a Xeon E5-2430, 6 cores, 2.20GHz, 48GB, under Linux Ubuntu 12.04.2 LTS. The Indexing Speed Factor (ISF), Searching Speed Factor (SSF) and Peak Memory Usage for indexing and searching (PMU_i and PMU_s) as described in [5] are almost the same for all runs (the differences between different runs stand only in the AM used). Their average values are ISF =0.81, SSF=$1.2*10^{-5}s^{-1}$, PMU_i=2203MB, PMU_s=197MB.

3. CONCLUSIONS

We have approached STD with a two step process. Single or multilingual 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 show that by training with multiple languages the accuracy of the detection is increased, however the quantity of the data used for training is insufficient for training such a large phoneme set. The searching algorithm works better for query types 1 and 2 and slightly worse for query type 3 where the words' order may be inverted.

4. REFERENCES

[1] X. Anguera, L.J. Rodriguez-Fuentes, I Szöke, A. Buzo and F. Metze, "Query by Example Search on Speech at Mediaeval 2014", in Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain, October 16-17.

[2] H. Andi Buzo, Horia Cucu, Iris Molnar, Bogdan Ionescu and Corneliu Burileanu, "SpeeD@MediaEval 2013 : A Phone Recognition Approach to Spoken Term Detection", in Proc. Mediaeval 2013 Workshop, Barcelona, Spain, 2013.

[3] J.S. Garofolo, et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Linguistic Data Consortium, Philadelphia, 1993.

[4] http://www.langsci.ucl.ac.uk/ipa/

[5] L.-J. Rodriguez-Fuentes and M. Penagarikano. MediaEval 2013 Spoken Web Search Task: System Performance Measures. Technical report, GTTS, UPV/EHU, May 2013.