A Study of the Influence of Speech Type on Automatic Language Recognition Performance

Alejandro Abejón, Doroteo T. Toledano, Danilo Spada, Victor González, Daniel Hernández
ATVS Escuela Politécnica Superior, Universidad Autónoma de Madrid
E-mail: {alejandro.abejon, doroteo.torre, danilo.spada, victorm.gonzalez, d.hernandezlopez}@uam.es

Abstract

Automatic language recognition on spontaneous speech has experienced a rapid development in the last few years. This development has been in part due to the competitive technological Language Recognition Evaluations (LRE) organized by the National Institute of Standards and Technology (NIST). Until now, the need to have clearly defined and consistent evaluations has kept some real-life application issues out of these evaluations. In particular, all past NIST LREs have used exclusively conversational telephone speech (CTS) for development and test. Fortunately this has changed in the current NIST LRE since it includes also broadcast speech. However, for testing only the telephone speech found in broadcast data will be used. In real-life applications, there could be several more types of speech and systems could be forced to use a mix of different types of data for training and development and recognition. In this article, we have defined a test-bed including several types of speech data and have analyzed how a typical language recognition system works using different types of speech, and also a combination of different types of speech, for training and testing.

1. Introduction

In the last few years we have started to use regularly automatic language technology operating on text, but automatic language recognition on speech is still much more complex and limited in performance. This technology has experienced a rapid development in the last years that has been in part due to the competitive technological Language Recognition Evaluation (LRE) organized by the National Institute of Standards and Tecnology (NIST). However, these evaluations have some limitations to keep the evaluation under control. One of these limitations is that NIST LREs have always been based on conversational telephone speech (CTS), while in real life applications language recognition systems typically are required to work with other types of data or be trained on different types of speech. The last NIST LRE, currently in progress, will start to look into other types of speech. In particular, it provides training speech taken from the media (i.e. Broadcast Speech, BS), but for testing it only uses parts of BS that are detected and audited to be telephone speech. This paper tries to provide an starting point to analyze the problem of how using different types of speech for training and test influences language recognition performance, how the train-test speech type mismatch influences performance and how mixing different speech types for training can alleviate the problem.

This article continues with a description of the language recognition technique used (section 2), a description of the experimental set-up including the corpora used in the experiments (section 3), the results obtained (section 4) and ends with a discussion and a conclusion.

2. Language Recognition System Used

Since our main interest in this work was the analysis of the influence of the speech type on language recognition performance we decided to use a relatively simple and well known system: a Paralell Phonetic Recognition followed by Language Modeling (PPRLM) system. Most state-of-the-art systems today tend to use a combination of several acoustic and phonotactic systems for increased performance. Here we decided to use a simpler phonotactic system due to the long tradition of the PPRLM systems in Language Recognition and to avoid the difficulties in determining which system causes the decrease in performance. In any case, our interest with this study is not comparing systems but comparing results on different databases and types of speech to assess its influence in language recognition.

Phonotactic system, such as the PPRLM system, try to model the sequences of phonemes that are characteristic of a particular language. An analogy with language recognition in text would be trying to recognize a language by looking at the most frequent sequences of letters. All phonotactic language recognition systems use a first block known as Phonetic Recognizer (PR) which transforms speech into a sequence of phonetic labels. Specifically, a Paralell PRLM, like our system, uses a set of phonetic recognizers to improve its performance. Our phonetic recognizers are relatively simple context and speaker independent phonetic recognizers based on Hidden Markov Models (HMMs).

Once a PR is available, the phonetic decodings it produces can be used in different ways for language recognition. The most classical approach is to use statistical Language Modelling (LM) techniques to model the frequencies of phones and phone sequences (n-grams) for each particular language. The combination of a single PR and LM gives the PRLM approach. If several PRs are used we obtain the PPRLM approach (Zissman, 1996). The PPRLM approach has dominated the field of language recognition for years and is still one key subsystem of state-of-the-art language recognition systems.

Our system uses 7 Phone Recognizers (PRs) in 7 different languages, which gives a total of 7 different PRLM systems that were fused. Six PRs were trained on six SpeechDat-like corpora, each of which contains over 10

hours of training material covering hundreds of different speakers in a single language. The languages of these PRs and the corresponding corpora used are English (with the corpus with ELDA catalogue number S0011), German (S0051), French (S0185), Arabic (S0183 + S0184), Basque (S0152) and Russian (S0099). We have also included a 7th PR in Spanish trained on Albayzin (Moreno et al., 1993) downsampled to 8 kHz, which contains about 4 hours of speech for training. All these PRs are based on Hidden Markov Models (HMMs) trained using HTK and used for recognition with SPHINX. The phonetic HMMs are three-state left-to-right models with no skips, being the output pdf of each state modeled as a weighted mixture of 20 Gaussians. The acoustic processing is based on 13 Mel Frequency Cepstral Coefficients (MFCCs) (including C0) and velocities and accelerations for a total of 39 components, computing a feature vector each 10ms and performing Cepstral Mean Normalization (CMN).

Each of the seven different PRLM subsystems is based on the following steps. First a voice activity detector segments the test utterance into speech and non-speech segments, which optimizes the whole process (Toledano et al., 2007). The speech segments are recognized with one PR. The recognized phonetic sequence is used to estimate 3-grams and these used to compute the probabilities. The scores produced by each of the seven different PRLM subsystems are normalized using T-Norm and fused using sum fusion to produce the final score, which is again T-Normalized.

3. Experimental Set-up

The main goal of these experiments was to analyze the impact of the database and type of speech on the performance of language recognition systems, the adequate definition of training and test data covering different types of speech data is crucial.

We limited number of language of recognition to five language: Arabic, English, French, German and Russian. For these 5 languages we have data of three different types that we have considered for our experiments:

- *Conversational Telephone Speech* (CTS) data: this data is from the CallFriend. This was used in past NIST evaluations. CallFriend's file contains recordings of 5 to 30 minutes of speech from different speakers talking to other human. For Russian we have used the RusTen corpus, which is very similar to CallFriend in its contents.
- *SpeechDat* (SD) data: This is data coming from SpeechDat corpora. Although this data is also telephone speech it is not Conversational Telephone Speech (CTS). It could more accurately described as prompted speech. A machine asks humans to utter some short utterances over the phone and their utterances are recorded. This is the type of speech that is usually expected in an automatic Interactive Voice Response (IVR) system.
- *Broadcast News* (BN) data: This is data recorded from several TV and radio stations (mainly from news). The speech is somewhat spontaneous (although most of it is read) and has very high quality.

To normalize the amount of data used for training and testing for each type of data we decided to use only the minimum amount of speech per language available for each type of data. For CTS and SD data we have over 20 hours of speech, but for BN the amount of speech is more limited (only 6 hours). We have subdivided these amounts of speech into two subsets, one for training the language models and other for testing the language recognition performance. The amount of speech used for each language and speech type.

CORPORA	SD	CTS	BN
TRAIN	16 h	20 h	4 h
TEST	6 h	7 h	2 h
TOTAL	24 h	27 h	6 h

Table 1: Amount of net speech per language in hours in training and test subsets for the different speech corpora used in the experiments: SpeechDat (SD), Conversational Telephone Speech (CTS) and Broadcast News (BN).

Following the convention used in NIST LREs we have normalized the duration of the test segments used for evaluating the language recognition systems. In particular we have made all test segments contain about 30s of net speech, which is equivalent to the main condition used in NIST LREs.

4. Results

In this section we present the experimental result obtained, which stress the important influence of type of speech in automatic language recognition performance.

The first experiment analyzes the influence of mismatch, when training and testing data are of different type. This is a relatively common situation in real-life applications of language recognition systems that appears typically when you try to recognize a speech of a particular type for a language for which you don't have training speech of that particular type. A particular example of this type of mismatch (where the target speech type is CTS and training is with BN speech) is currently under evaluation in NIST LRE 09.

For the experiment we have trained the language models for the 5 language used in our data set using each of the three different types of speech data considered: CTS, SD and BN. Then we have evaluated performance using these models with the test data of each of the three different speech types considered.

Table 2 shows the Equal Error Rates (EER) obtained for each of the 9 different combinations of training and test speech types for each of the 5 languages considered, as well as the EERs averaged for all languages. For each language to detect and test speech type the best result obtained for the different types of training speech is highlighted in boldface.

LANGUAGE	TRAIN \ TEST	SD	CTS	BN
RUSSIAN	SD	**2.53**	40.56	30.29
	CTS	31.08	**2.78**	24.52
	BN	20.48	14.44	**8.65**
ARABIC	SD	**5.83**	28.00	36.07
	CTS	32.76	**6.58**	2.74
	BN	33.18	12.67	**1.37**
GERMAN	SD	**8.10**	23.67	28.79
	CTS	11.16	**8.88**	23.74
	BN	24.07	31.00	**8.59**
ENGLISH	SD	**0.90**	34.67	41.21
	CTS	29.88	**6.25**	7.04
	BN	21.66	20.00	**5.03**
FRENCH	SD	**2.34**	40.00	**11.27**
	CTS	35.94	**6.58**	45.10
	BN	19.40	17.00	26.47
Average per Language	SD	**3.94**	33.38	29.53
	CTS	28.16	**6.21**	20.63
	BN	23.76	19.02	**10.02**

Table 2: Language Recognition Results (EER in percentage) for different combinations of training and test speech corpora for different languages.

As expected, the best result is always obtained when the speech types used for training and test match. The only exception to this result is the result for French and test material of type BN. The only exception to this is the case of French and BN test data for which SD training outperforms BN training possibly due to the reduced amount of BN training or a dialect mismatch. Averaged results show clearly that mismatch between training and testing speech type has an important influence on performance, making the ERR at least double (and sometimes even multiply by 6). This gives an idea of the important influence of the speech type in language recognition performance.

Once we have detected this problem, we have experimented with perhaps the most obvious way to try to reduce the influence of this problem: using multi-condition training (i.e. training our models with different types of speech to try to make them more robust against speech type mismatch). Table 3 shows results obtained by training the language recognition models with the three possible combinations of 2 of the 3 types of speech considered, as well as with the three types considered (ALL). In this table we have highlighted the worst result (in EER) for each language and test condition. As can be observed, the worse results tend to be obtained when the test speech type is not considered for training, which indicates that multi-condition training seems to be a valid procedure to alleviate the problem of speech type mismatch in language recognition. Furthermore, the combination of the three different types of speech for training never gets the worse results. In fact, comparing the average results in Table 3 obtained training with the three different types of speech with those of Table 3 training with the matched type of speech it can be

LANGUAGE	TRAIN \ TEST	SD	CTS	BN
RUSSIAN	SD+CTS	1.53	10.00	14.42
	SD+BN	1.62	**34.44**	**16.83**
	CTS+BN	**19.54**	2.78	16.34
	ALL	1.34	9.44	12.5
ARABIC	SD+CTS	5.83	9.00	**9.59**
	SD+BN	5.83	**18.33**	2.74
	CTS+BN	**26.43**	4.33	1.83
	ALL	6,40	6,67	2,74
GERMAN	SD+CTS	5.40	**17.67**	**17.17**
	SD+BN	6.59	**17.67**	8.59
	CTS+BN	**8.81**	11.00	5.56
	ALL	5,66	13,00	12,12
ENGLISH	SD+CTS	1.35	13.00	**19.10**
	SD+BN	0.96	**24.00**	6.53
	CTS+BN	**14.10**	6.33	3.02
	ALL	1,07	13,00	7,03
FRENCH	SD+CTS	2.37	17.33	12.75
	SD+BN	2.74	**35.67**	9.80
	CTS+BN	**20.67**	6.67	**37.75**
	ALL	2,60	14.67	11.27
Average per Language	SD+CTS	3.30	13.40	**14.61**
	SD+BN	3.55	**26.02**	8.90
	CTS+BN	**17.91**	6.22	12.90
	ALL	3,41	11,36	9,13

Table 3: Multi-condition training Language Recognition Results (EER in percentage) for different combinations of two types of speech for training and one type of test speech for different languages.

observed that result are similar. In average for two of three test speech types (BN and SD) results are actually better than the results obtained with matched training and test (Table 2). For the remaining type of speech (CTS) the average EER for the matched condition is 6,27% and it degrades to 11,36 % for the multi-condition training with all speech types. We consider that one possible reason for this exception is that CTS is the most abundant type of speech in our experiment, so that the influence of other types of speech is less important than in the case of BN and SD. So we can conclude that multi-condition training seems to work relatively well for language recognition and it is desirable to use it when available. Multi-condition training seems to work similar to matched condition for some speech types. However, there are speech types that seem to be more difficult to model in combination, such as CTS, so we see that multi-condition training is not the best solution to the data type mismatch and further research is definitely desirable.We have also analyzed the behavior of score distributions in this experiment. We have analyzed separately the statistical distribution of the scores produced for target trials (i.e. comparison of a test segment and a model of the same language) and non-target trials (i.e. comparison of a test segment and a model of different languages) for the different sets of models trained, the

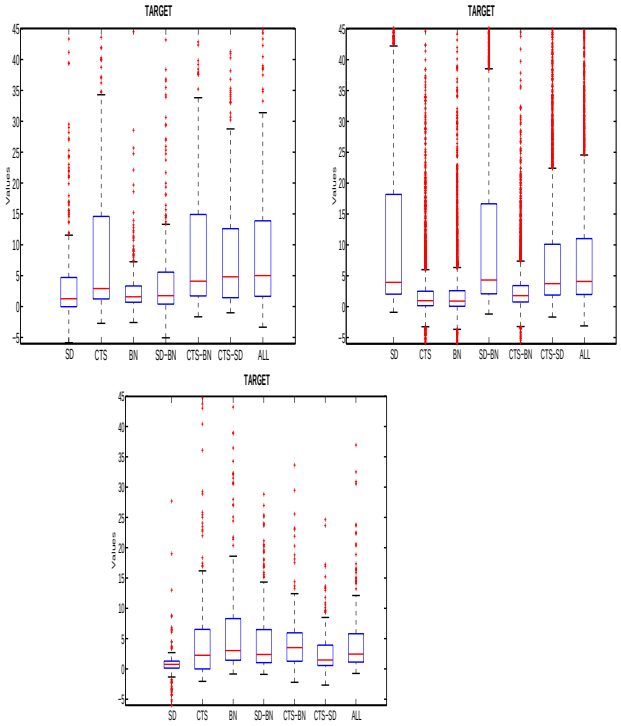

Figure 1: Distribution of target scores for CTS test data and the different models trained with the different types of data and combinations of them (top left) . Distribution of target scores for SD test data and the different models trained with the different types of data and combinations of them (top right). Distribution of target scores for BN test data and the different models trained with the different types of data and combinations of them.(bottom).

different languages and the different train and test speech types. Non-target scores (not included in Fig 1 due to space restrictions) tend to be similar in mean, but usually tend to have larger variance for the case in which the type of test segment is not considered in training. Target scores, on the other hand, present variation in the means, which tend to be higher when type of the test segment is considered in training (see Fig 1). The variances of target scores behave opposite to those of the non-target scores, i.e, these variances tend to be higher for cases in which the type of speech of the test is considered during training. This difference in score distribution suggest that appropriate score calibration (provided adequate data is available) could also alleviate the problem of speech type mismatch.

5. Conclusion

This study has shown that there is a strong influence of the speech type on language recognition performance. In particular mismatch between training and testing data speech type is able to spectacularly degrade system performance. This effect was not taken into account in NIST evaluations so far (NIST LRE 09, however, will start to deal with this type of problem), but in real-life applications this issue is a crucial factor.

We have shown that the score distributions produced by models trained on different speech types tend to be different both in terms of means and variances. This makes very dangerous to use together models trained on different types of speech (we have not mixed models trained on different types of speech in the experiments presented here, but attempts to do so have been very frustrating) as well as to allow mismatch between training and test. Our experiments show that the Equal Error Rates can be multiplied by a factor between 2 and 6 when moving from a matched to a mismatched condition.

We have experimented with one way to alleviate the problem of speech type mismatch: multi-condition training. This method has been quite successful to deal with different types of testing speech. However, multi-condition training requires to have access to training speech of all possible types, which is normally impossible in real-life applications.

Hence, we have measured the effect of speech type mismatch in language recognition, as well as the improvements that can be achieved with multi-style training. In future work, we will try to develop more powerful techniques to deal with the problem of speech type mismatch.

6. References

CallFriend corpora (1996) available for purchase from the Linguistic Data Consortium (LDC) on http://www.ldc.upenn.edu/Catalog/byType.jsp#speech, catalog codes: LDC96S46 to LDC96S60.

ELDA Catalogue, European Language Distribution Association, www.elda.org.

Hidden Markov Model ToolKit, available on http://htk.eng.cam.ac.uk/ .

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. and Nadeu, C. (1993). ALBAYZÍN Speech Database: Design of the Phonetic Corpus, in *Proceedings of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH)*. Berlin, Germany, 21-23, Vol. 1. pp. 175-178.

Sphinx toolkit. available on http://cmusphinx. sourceforge.net/html/cmusphinx.php.

Toledano, D.T., González-Domínguez, J., Abejón-González, A., Spada, D., Mateos-García, I. and González-Rodríguez, J. (2007). Improved Language Recognition Using Better Phonetic Decoders And Fusion With MFCC, And SDC Features, in *Proceedings INTERSPEECH*, pp 194-197.

Zissman, M.A. (1996) Comparison of four approaches to automatic language identification of telephone speech., *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 1, pp. 31–44,