Acquisition of Telephone Data from Radio Broadcasts with Applications to Language Recognition

Technical Report

Oldřich Plchot, Valiantsina Hubeika, Lukáš Burget, Petr Schwarz, Pavel Matějka and Jan "Honza" Černocký

Speech@FIT, Brno University of Technology, Czech Republic,
{iplchot,ihubeika,burget,schwarzp,matejkap,cernocky}@fit.vutbr.cz

January 2009

Summary

Current LID systems have difficulties in dealing with languages with insufficient or small amount of training data available. This issue concerns not only exotic languages with small number of native speakers, but also languages like Thai with 65 milion native speakers.

We aim to develop techniques, that will allow us to automatically obtain training data for these troubled languages and use them in Language Recognition systems. As the present LRE systems are trained and evaluated on Continuous Telephone Speech (CTS), the task will be to obtain speech samples, that went through the telephone channel. This task leads us to developing an automatic system, which obtains recordings from public broadcasts and automatically detects telephone calls that are consequently used for training. The system was implemented and used for building the data sets which were used for subsequent experiments.

In order to use the data obtained from broadcasts we have to cope with several issues related to this data. The first problem is channel compensation, as the data comes not only through telephone channel, but also through wideband broadcast. The second problem is that the telephone calls into broadcasts are usually less spontaneous than data commonly used for current systems.

We have conducted several experiments using both CTS and broadcast data to uncover possible problems, which can arise when using this type of data in training or evaluating current LRE systems. The results of these initial experiments show that if the broadcast data only are used for training and standard telephone data for testing, the performance of such system is worse, than the performance of standard LRE systems trained and tested on CTS. Further experiments with compensation for the distortion created by broadcast channel should be conducted to better match the target CTS data and improve the performance.

The experiments also show, that if the broadcast data are used both for training and testing the system, the results are very good. This can indicate, that the information about channel is very strong in these broadcast data and that the systems are learning this information and it heavily affects the final recognition.

Cooperation with Linguistic Data Consortium on creating a broadcast database was part of this work. We used the developed systems to provide pre-labeling of broadcast data, see Appendix B.

Acknowledgements

This work was supported by US Air Force European Office of Aerospace Research & Development (EOARD) under Grant No. 083066.

Contents

Chapter 1

Introduction

We introduce a process of automatic acquisition of speech data from the various media sources for the language identification task. The last editions of NIST Language Recognition (LRE) evaluations have shown that both acoustic and phonotactic approaches have reached a certain maturity level in both modeling of target languages and dealing with the influences of different channels. However we are still facing the common problem: the lack of training data. There is no good or large enough database of training data for many languages including even languages like Thai, which is spoken by 65 million speakers. Also, there is an increasing demand to recognize languages from smaller and less populous regions (many of them relevant for security of defense domain). For some of these languages, no standard speech resources exist.

This work aims at solving this problem using the data acquired from public sources, such as satellite and Internet TVs and radios, which contain conversational speech or telephone calls. This approach can provide us with large amount of data that we will use to conduct experiments, which will help to answer the question whether these data can replace or augment standard conversational telephone speech (CTS) data. The results will also show that if we had no standard CTS training data, these data obtained from broadcasts can be used to process the languages that we were unable to recognize due to absence of the training data.

First, the obtained data has to be preprocessed in order to acquire clean speech segments or individual phone calls. The task is to examine the obtained telephone calls by training and evaluating the systems on languages for which we have both CTS and broadcast data. The results of the experiments will show, how the systems perform, when the CTS or broadcast data are used for training or testing.

The main challenge is channel compensation, as the obtained data are acoustically very different from the conversational telephone speech (CTS) commonly used in LRE. Broadcast data contain a great deal of unspontaneous speech as well. Further task is to explore how unspontaneous speech affects current LRE systems (which are supposed to be trained on spontaneous data). The notion of channel compensation will therefore have to be extended to cope with these factors.

We have done experiments on Dari, English, French, Hindi, Korean, Mandarin, Spanish and Vietnamese languages, because these languages are the intersection of languages we obtained from broadcast sources and the languages present in standard databases available.

Chapter 2

Data Acquisition Principles

There is unlimited source of speech data available from the broadcast media. We can acquire data from several sources, each of which has different channel parameters, quality and number of available languages. The list of available sources in a standard industrialized country (such as the Czech Republic) is shown in Table 2.1 [1].

All of the listed sources except Internet radios are geographically dependent regarding location. The quality of different Internet sources varies a lot and it is important to carefully choose them. We have used an archive[1] of Voice of America Internet radio to obtain data for all languages.

This particular data of VoA were obtained in MP3 format, bitrate is 24 Kbit/s, sampling rate 22,050 Hz, 16 bit encoding, mono. Original media data include a great portion of music and speech with the music in background. We have to deal with this problem and select only clean speech segments. Also we should deal with the problem of a low speaker variability in the obtained data, for instance as it is common in news programmes, which are moderated by the same speaker. So far, we have not investigated into this problem and used only telephone calls in broadcasts, where speaker variability should be sufficient.

2.1 Detecting Phone Calls

Our phone call detector is based on the fact that a telephone channel acts like a bandpass filter, which passes energy between approximately 400 Hz and 3.4 KHz. On the other hand, regular wideband speech contains significant energy up to around 5 KHz. Common media sources like satellite radio or Internet radios are usually sampled at 22 kHz so it supports this bandwidth, which means that if we place a phone call into the regular radio transmission, we will see a significant change in the spectrum (Figure 2.1).

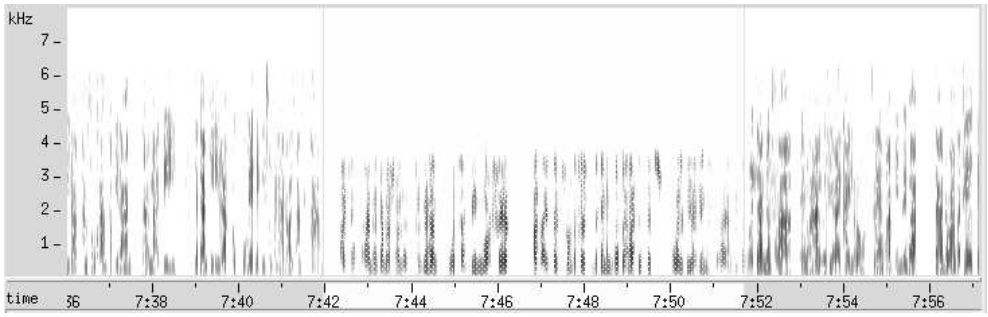

Figure 2.1: Phone Call in a Radio Broadcast.

[1]FTP server 8475.ftp.storage.akadns.net directory /mp3/voa

Table 2.1: Overview of different channels. DVB stands for Digital Video Broadcasting - Terrestrial, Cable and Satellite. By parallel recording we mean the possibility of acquiring more broadcasts simultaneously using one recording device (i.e. one DVB-S receiver).

	Inet. radio	**DVB-T**	**DVB-C**	**DVB-S**	**Analog**
Languages	approx. 100	1 - 3	approx. 5	20 - 30	3 - 5
Quality	variable	good	good	good	bad
Parallel recording	yes	yes	yes	yes	no

For the detection, we first resample the signal to commonly used 16 kHz. The signal is divided into frames of 512 samples with no overlap and Fourier spectrum is computed for each frame. To detect boundary between wideband and telephone speech, we concentrate on the frequency range between 2350 and 4600 Hz. The power spectral density (PSD) in this range was used (see Figure 2.2). At first, the PSD was normalized to zero mean and unit variance. Then values in the first half (from 2350 to 3475 Hz) and values in the second half (from 3475 to 4600 Hz) of the PSD were summed. A ratio between these two sums was compared with a threshold and the decision was made. If the ratio is higher than selected threshold, there is more energy in lower frequencies and we considered the segment a telephone call speech. For the block diagram of this process see Figure 2.3.

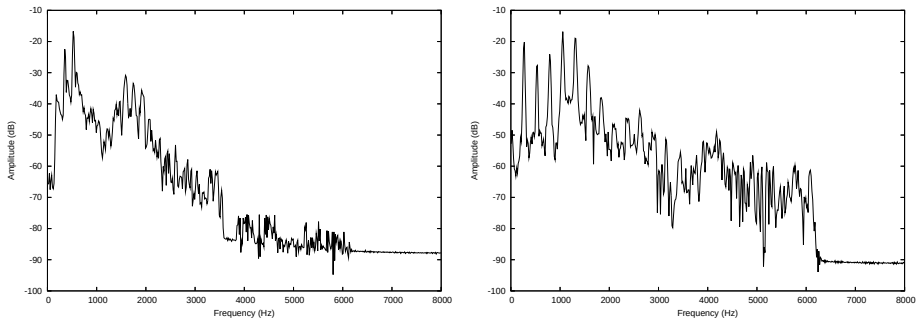

Figure 2.2: Power Spectral Density of telephone call in the broadcast (left figure) and wideband speech (right figure).

2.2 Detecting Wideband Speech Segments

Recordings obtained from media broadcasts contain great deal of music, speech with music in the background or other nonspeech sounds. The task is to detect clean speech segments which can be used in language recognition or possibly in the other applications.

The detection is done by estimating frame by frame likelihoods, of classes *speech* and *other* (non-speech). GMM models were used to estimate these likelihoods. These models contain 1024 Gaussians and were trained on 12.7 hours of speech and 18.7 hours of nonspeech wideband data. MFCC coefficients with deltas and double deltas were used as features for training. These data (containing several languages) were obtained from Linguistic Data Consortium and were manually annotated for these two classes.

Once we obtain frame by frame log-likelihoods for each class, we filter them using simple median filter[2] and subtract these two sets of values. The resulting log-likelihood ratios are averaged over 100 frames and compared to empirically set thresholds. Depending on the threshold, we decide whether we

[2]Window size of this median filter is 5.

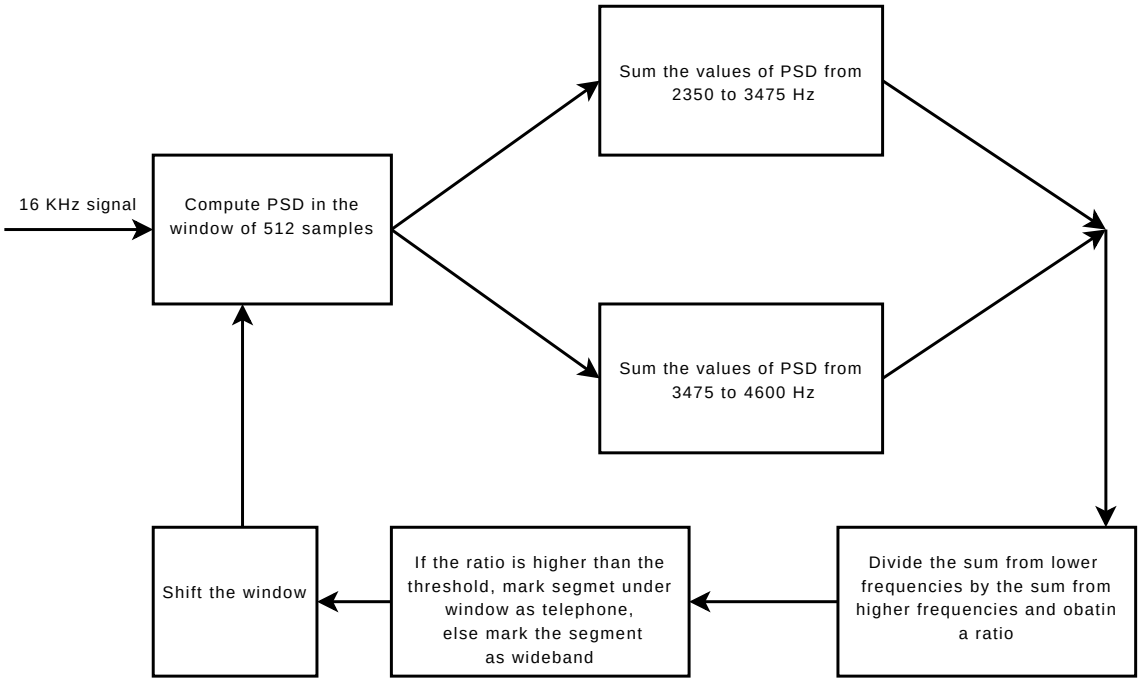

Figure 2.3: Block diagram of detecting telephone calls in the wideband signal.

are in the speech segment or nonspeech segment or whether we are not sure (segments to be checked by human annotator). For the block diagram of this process see Figure 2.4.

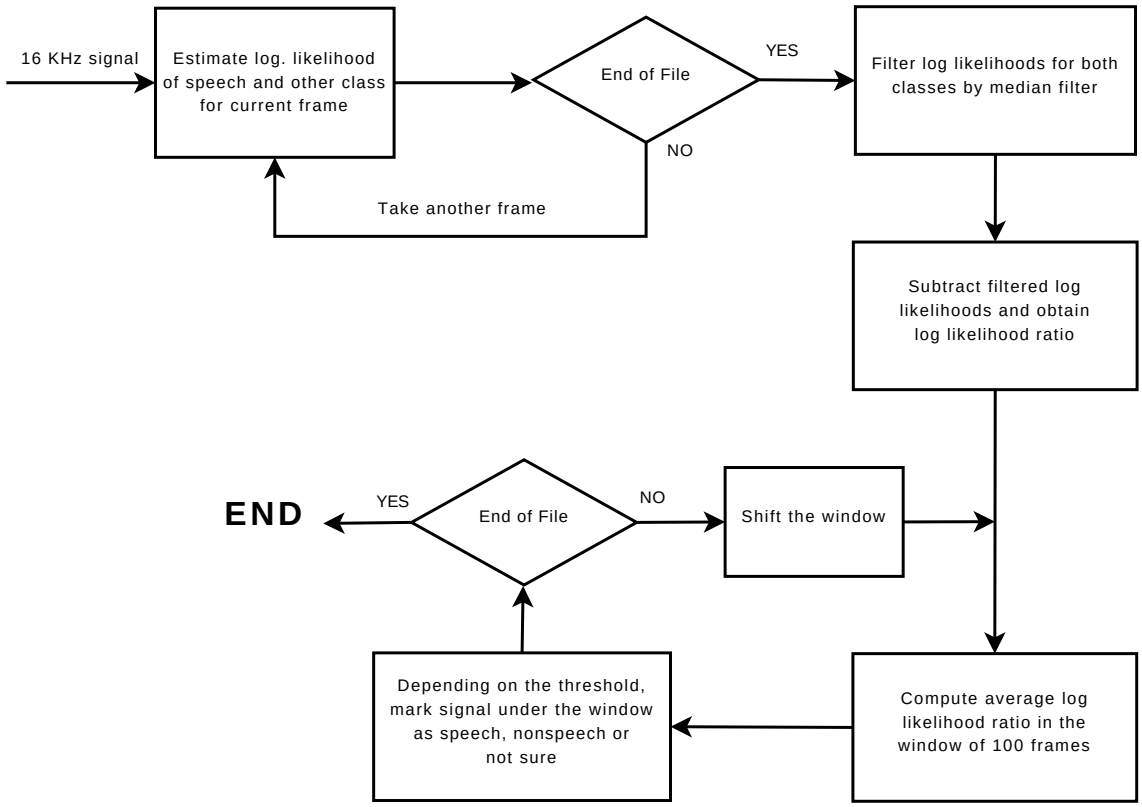

Figure 2.4: Block diagram of detecting speech and nonspeech segments in the wideband signal.

Chapter 3

Training and Test Sets

In order to compare, how our LRE systems perform when using broadcast and standard CTS data, we created a data set from broadcast data. We selected eight languages[1] from the Voice of America ftp archive. We have chosen these particular languages, because we have the data for these languages present in CallFriend, NIST LRE 2003 and NIST LRE2007 databases. In order to create reasonably robust experiment, we have chosen these languages even if we expected problems with French and Dari language: the French language in the Voice of America archive is recorded in the Africa region and therefore the obtained samples can substantially differ from the utterances spoken by native French speakers in our CTS databases. The *Dari* language was chosen, because this language is very close to the *Farsi* language which is present in CallFriend, NIST LRE 2003 and NIST LRE 2007 databases. We decided to relabel **Farsi to Dari** in those databases for the purpose of the experiments.

Additionally, we expect, that the people calling into the Voice of America broadcasts *speak the same language* as the language label denoting particular recording of broadcast. We did not have resources to manually check all data, so errors can occur in labeling of the training and test data. We have to keep in mind all of these compromises we have made when analyzing the results of the experiments.

3.1 Telephone Call Segments

We decided to select only telephone calls which are present in the Voice of America broadcasts, because we believe these data will be affected by passing through the telephone channel and will better match our CTS data. First, our phone call detector was used to *detect phone call segments* in the wideband data. The telephone call into broadcast can be interrupted by a moderator and we want to reconstruct the call from the segments of the calling person. The postprocessing of this detection was made in order to obtain these reconstructed segments.

For the purpose of the postprocessing of label file created by phone detector, an algorithm which marks particular phone segments as `phonecall1, phonecall2` ... was designed. This algorithm marks individual phone call segments in order to join them into longer segments. The algorithm accepts segments which are longer than *10 seconds*, because our phone call detector makes a lot of short segments, which are more likely to contain some wideband portion. Phone call segments are assigned the same label until there is a maximum *120 seconds* of wideband segment between them. When the wideband segment between phone calls is longer than 240 seconds, the next phone segments will be assigned new label (e.g. `phonecall2`).

When the label file created by the telephone detector is processed by the algorithm explained above, we cut and join the segments with the same label. Speech@FIT phone recognizer [2], [3] was used to determine the pause in the speech at the borders of each segment and these time stamps were

[1] Dari, English, French, Hindi, Korean, Mandarin, Spanish and Vietnamese

Table 3.1: Training data in hours after segmentation for each language.

Language	CallFriend	Broadcast
Dari/Farsi	21.2	6
English	39.8	6
French	21.5	6
Hindi	19.6	6
Korean	18.4	6
Mandarin	41.7	6
Spanish	43.8	6
Vietnamese	20.6	6

Table 3.2: Number of 30 second test segments for each language.

Language	NIST 2003	NIST 2007	Broadcast
Dari/Farsi	80	88	150
English	240	266	150
French	80	80	150
Hindi	80	268	150
Korean	80	108	150
Mandarin	80	496	150
Spanish	80	256	150
Vietnamese	80	168	150

used to cut the segments out of the original recordings. Then the cut segments with the same label were concatenated into one file to obtain the reconstructed telephone call.

Using this approach, we obtain significantly smaller number of telephone segments than we would get taking directly the output of the telephone detector. The benefit is that the segments contain less wideband caused by errors in detecting the phone calls and the speaker variability is increased, because we have less segments with the same speaker. On the other hand, it is possible, that the final segments contain more different speakers.

3.2 Broadcast Data Sets

Using the procedure explained above, we created *broadcast test set*, selecting 150 segments for each language. Each selected segment was cut out from the detected telephone call in such way, that it contained *30 seconds* of speech. Our phoneme recognizer was used to determine the length of speech.

Broadcast training set was created by taking the merged phone call segments[2] until we reached the limit of six hours of speech per language.

3.3 CTS Data Sets

CTS test sets were created by taking subsets of NIST LRE 2003 [4] and 2007 [5] evaluation data. Only *30 second* segments were used. Training set was created by taking subset of languages from CallFriend database. All data sets are listed in tables 3.1 and 3.2.

[2]Described in section 3.1

Chapter 4

Experiments

We performed experiments both with phonotactic and acoustic systems. With both systems, we tested several techniques to improve the performance to show in which direction the development of LRE systems using data obtained from broadcasts together with standard CTS data should continue. The results are evaluated using standard metrics: Detection Error Tradeoff (DET) curve, Decision Cost Function (DCF) and Equal Error Rate (EER) [5]. All experiments were done on *30 second segments*. We present results of phonotactic and acoustic systems derived from our systems submitted to NIST LRE 2007 evaluation [6, 7].

4.1 Phonotactic systems

The first phonotactic system [6, 7] is based on string output of our Hungarian phoneme recognizer. The second phonotactic system [6, 7] is based on lattice output of our Hungarian phoneme recognizer. The phoneme recognizer is based on hybrid ANN/HMM approach, where artificial neural networks (ANN) are used to estimate posterior probabilities of phonemes from Mel filter bank log energies using the context of 310ms around the current frame [3]. Trigram language models were trained on CallFriend database for CTS phonotactic system and for broadcast phonotactic system, the language models were trained on broadcast training set. Linear back-end calibration [8] was applied on the obtained scores. Calibration of scores was done on the test set, which may lead to overoptimistic results, but according to our experience, the results for properly trained calibration will not differ much. Both CTS and broadcast systems were evaluated against all test sets.

4.1.1 Results of Phonotactic Systems

The results are listed in tables 4.1 and 4.2. Phonotactic system based on string output was outperformed by the phonotactic system with lattices in all cases.

Table 4.1: Phonotactic systems based on string output - pooled EER

		TEST		
TRAIN		NIST 2003	NIST 2007	Broadcast
	CTS	1.781	9.072	6.583
	Broadcast	11.949	18.593	1.416

Table 4.2: Phonotactic systems based on lattice output - pooled EER

		TEST		
TRAIN		NIST 2003	NIST 2007	Broadcast
	CTS	0.900	6.995	5.232
	Broadcast	8.958	15.215	1.398

4.2 Acoustic systems

Our acoustic systems are built on the experience with GMM modeling for speaker recognition [9] which follows conventional Universal Background Model-Gaussian Mixture Modeling (UBM-GMM) paradigm [10] and employs number of techniques that have previously proved to improve GMM system performance [11]. This system was chosen because it can easily compensate for the channel distortion.

Table 3.1 lists the corpora used to train our systems. CTS system was trained on CallFriend database and broadcast system was trained on our broadcast database.

Our systems use the popular shifted-delta-cepstra (SDC) [12] feature extraction, where 7 MFCC coefficients (including coefficient C0) are concatenated with SDC 7-1-3-7, which totals in 56 coefficients per frame. Vocal-tract length normalization (VTLN) [13] performs simple speaker adaptation. VTLN warping factors are estimated using single GMM (512 Gaussians), ML-trained on the whole CallFriend database (using all the languages). The model was trained in standard speaker adaptive training (SAT) fashion in four iterations of alternately re-estimating the model parameters and the warping factors for the training data. Each language model is obtained by traditional *relevance MAP* adaptation [14] of UBM using enrollment conversation. Only means are adapted.

In verification phase, standard Top-N Expected Log Likelihood Ratio (ELLR) scoring [14] is used to obtain verification score, where N is set to 10. However, for each trial, both language model and UBM are adapted to channel of test conversation using simple eigenchannel adaptation [9] prior to computing the log likelihood ratio score.

Calibration of scores was done on the test set, which may lead to overoptimistic results, but according to to our experience, the results for properly trained calibration will not differ much. Both CTS and broadcast systems were evaluated against all test sets.

4.2.1 Results of Acoustic Systems

First, both systems were trained without channel compensation. Then, eigenchannel adaptation was applied. Two different matrices containing 50 eigenchannels were used. The first matrix was computed from broadcast training set. The second matrix was taken from our NIST LRE2007 system [6]. This matrix was trained on CTS databases.

We also experimented with training channel compensation using both CTS and data from broadcasts, hoping that the channel compensation will solve the mismatch between CTS and broadcasts. Especially we were hoping to improve the poor results when training on broadcasts and testing on CTS. However, so far we were not successful with such cross-condition channel compensation.

The results are listed in tables 4.3, 4.4 and 4.5.

Table 4.3: Acoustic systems without eigenchannel compensation - pooled EER

		TEST		
TRAIN		NIST 2003	NIST 2007	Broadcast
	CTS	3.407	8.807	8.261
	Broadcast	14.423	19.502	3.250

Table 4.4: Acoustic systems with eigenchannels trained on broadcast data - pooled EER

		TEST		
TRAIN		NIST 2003	NIST 2007	Broadcast
	CTS	1.145	5.644	8.250
	Broadcast	9.840	15.013	0.583

4.3 Discussion

The results of both acoustic and phonotactic systems were consistent. Phonotactic systems using lattices significantly outperform phonotactic systems based on string output in all test cases. See Appendix A for detailed results.

We expected that the acoustic systems outperform phonotactic systems, but only phonotactic system trained on CTS was outperformed by acoustic system trained on CTS with channel compensation trained on telephone data.

The results of acoustic systems prove that the individual samples are recorded over different channels, therefore application of eigenchannel adaptation [15] is crucial to compensate the channel distortion. In language detection task, channel variability may comprehend not only variability in the telephone channel or type of microphone, but also session or speaker variability.

Channel compensation trained on CTS is generally better. Broadcast data probably do not reflect the variations of channels.

The results of acoustic systems trained on broadcast data can imply, that the wideband channel added additional distortion to the obtained data, which affects the results obtained when testing against the CTS data. The decline in performance when testing against the CTS data can be also affected by different type of speech, that is usually present in the broadcasts. Speech in media broadcasts is usually less spontaneous. Speech in radio broadcasts in comparison with our CTS databases does not contain many hesitations, interruptions and is usually grammatically correct.

However the performance of systems trained on broadcast data and tested on CTS data is worse than the performance of systems trained and tested on CTS, the results show the similar trend over individual languages. This trend when EER is approximately two times higher except for the Dari and French language [1], can be observed on NIST 2007 test set (see figures 4.1 and 4.2), which consists of more difficult data for recognition.

[1] We expected problems for these languages, see section 3.

Table 4.5: Acoustic systems with eigenchannels trained on CTS data - pooled EER

		TEST		
TRAIN		NIST 2003	NIST 2007	Broadcast
	CTS	0.420	4.296	3.083
	Broadcast	9.222	14.290	0.922

When evaluating the acoustic system trained on broadcast data, we obtain excellent performance on broadcast data, which can indicate, that the system learned also the different channels of individual radio stations. This hypothesis has to be kept in mind when using broadcast data both for training and testing. Channel compensation trained on broadcasts even emphasizes this possible problem.

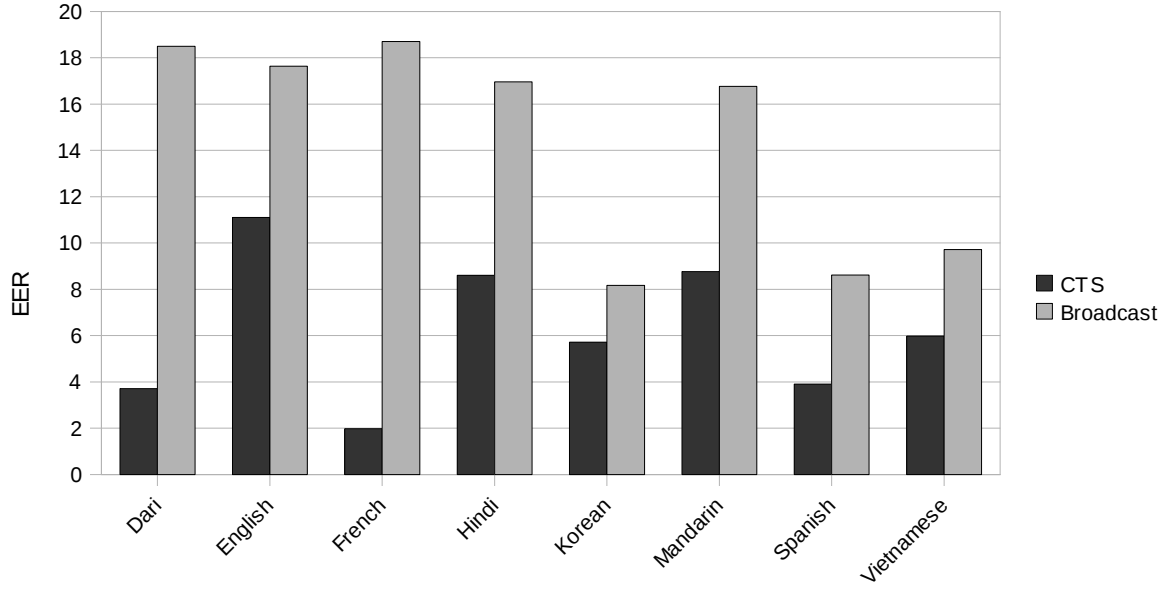

Figure 4.1: Equal Error Rate of individual languages for phonotactic systems based on lattices trained on CTS and broadcasts.

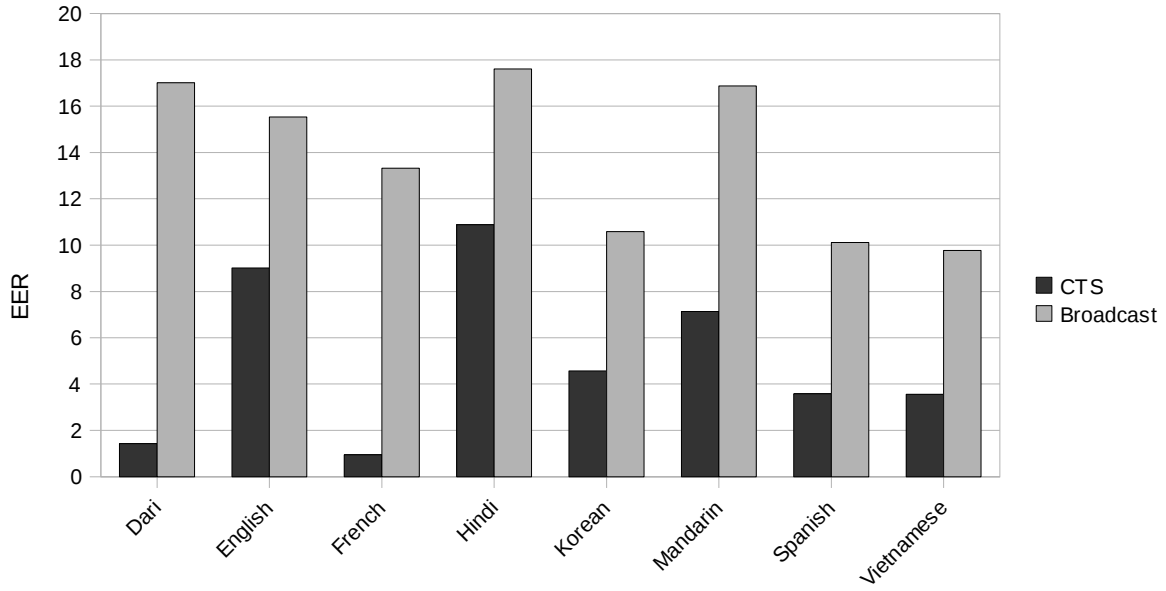

Figure 4.2: Equal Error Rate of individual languages for acoustic systems trained on CTS and broadcasts with channel compensation trained on broadcast data.

Chapter 5

Conclusions

We introduced a simple but promising approach of acquiring telephone data for LID. Experiments with selected languages using standard telephone data and telephone data acquired from broadcast were performed. Both phonotactic and acoustic approaches for recognition were investigated.

Obtained results show, that if systems trained on broadcast data are used to recognize CTS, the performance is significantly lower than it would be with the systems trained on target data. However, experiments with channel compensation techniques indicate, there is a possibility to improve the performance by investigating other compensation techniques to suppress the distortion caused by passing the telephone call through wideband channel. On the other hand, training the systems on CTS data and testing on broadcast data seems to be all right as the same trends are observed for the CTS based test sets.

Performed experiments show, that if broadcast data are used both for training and testing, the performance is excellent but if the CTS data are used to evaluate the system, the performances drop dramatically. This is probably because the systems trained and tested on broadcast data have learned some information about the channel of particular broadcast, especially if all samples of the same language come from one radio station, but this problem deserves further investigation. As soon a database exists, where one language comes from different broadcasts, the experiments should be made to verify this idea.

Results of the experiments also lead to a claim, that the broadcast data are "easier", as they contain mostly clean, prepared and grammatically correct speech. This idea is supported by the fact, that broadcast data were always (except the case when the channel compensation trained on broadcasts was used) recognized by systems trained on CTS data with better accuracy than NIST 2007 data which contain a lot of unclean speech.

It should be remembered, that the results of systems trained on broadcasts were obtained on automatically created databases without human annotator checking and several compromises were made, especially considering Farsi language as Dari and using French spoken in the African region. Also only *6 hours* of training data per language was used to train systems on broadcast data in comparison with average *28 hours* of training data per language for systems trained on CTS data.

The performance of the systems trained on broadcast data simulates a scenario, when no standard CTS training data are available and we need to detect a particular language. Although the results are significantly worse than ones we would get with CTS data for training, using the broadcast data can be the only option in such situation.

Appendix A

Detailed Results

Table A.1: Results of phonotactic system based on string output. System trained on CallFriend database.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	1.517	3.858	10.476
English	2.529	10.682	3.761
French	2.440	2.237	10.285
Hindi	2.142	10.202	6.238
Korean	0.208	6.142	5.000
Mandarin	0.952	11.166	4.000
Spanish	0.714	4.343	1.523
Vietnamese	0.684	7.314	5.619
Average	1.398	6.993	5.863
pooled minDET	1.700	7.627	6.261
pooled EER	1.781	7.736	6.583
pooled unweighted minDET	1.794	9.072	6.261
pooled unweighted EER	1.982	9.122	6.583

Table A.2: Results of phonotactic system based on string output. System trained on broadcast database.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	19.077	19.371	0.7142
English	11.666	19.698	2.666
French	13.839	20.968	1.285
Hindi	17.440	21.619	0.904
Korean	6.994	12.737	1.142
Mandarin	9.017	22.321	0.238
Spanish	5.000	10.200	0.666
Vietnamese	4.970	12.335	0.285
Average	11.000	17.406	0.988
pooled minDET	11.644	18.286	1.333
pooled EER	11.949	18.593	1.416
pooled unweighted minDET	12.035	18.796	1.333
pooled unweighted EER	12.250	19.122	1.416

Table A.3: Results of phonotactic system based on lattices. System trained on CallFriend database.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	0.744	3.707	6.714
English	1.726	11.108	3.761
French	0.803	1.976	7.761
Hindi	0.446	8.609	5.523
Korean	0.208	5.720	3.571
Mandarin	0.535	8.762	2.142
Spanish	0.148	3.904	1.857
Vietnamese	0.625	5.977	4.428
Average	0.654	6.221	4.470
pooled minDET	0.822	6.903	5.083
pooled EER	0.900	6.995	5.232
pooled unweighted minDET	0.866	7.769	5.083
pooled unweighted EER	0.875	7.836	5.232

Table A.4: Results of phonotactic system based on lattices. System trained on broadcast database.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	14.791	18.498	0.428
English	10.029	17.637	1.904
French	11.339	18.696	1.428
Hindi	12.559	16.958	0.666
Korean	3.839	8.172	1.142
Mandarin	5.446	16.762	0.333
Spanish	2.142	8.616	0.904
Vietnamese	2.886	9.717	0.047
Average	7.879	14.382	0.857
pooled minDET	8.697	15.017	1.220
pooled EER	8.958	15.215	1.398
pooled unweighted minDET	9.258	15.313	1.220
pooled unweighted EER	9.607	15.497	1.398

Table A.5: Results of acoustic system trained on CallFriend database without channel compensation.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	2.083	2.359	5.190
English	1.488	13.291	10.619
French	4.077	2.531	12.333
Hindi	4.255	15.340	13.238
Korean	2.291	7.788	6.238
Mandarin	2.559	10.153	5.666
Spanish	4.315	8.564	2.761
Vietnamese	1.160	3.661	4.190
Average	2.779	7.961	7.529
pooled minDET	3.277	8.670	8.184
pooled EER	3.407	8.807	8.261
pooled unweighted minDET	3.276	10.601	8.184
pooled unweighted EER	3.375	10.873	8.261

Table A.6: Results of acoustic system trained on broadcast database without channel compensation.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	21.577	23.920	1.142
English	9.375	22.564	3.761
French	16.220	20.010	3.666
Hindi	20.952	24.347	2.476
Korean	11.428	15.325	3.190
Mandarin	9.017	20.200	0.714
Spanish	10.000	14.083	1.428
Vietnamese	7.351	7.847	1.857
Average	13.240	18.537	2.279
pooled minDET	13.958	19.317	2.833
pooled EER	14.423	19.502	3.250
pooled unweighted minDET	13.357	20.133	2.833
pooled unweighted EER	13.633	20.350	3.250

Table A.7: Results of acoustic system trained on CallFriend database with channel compensation trained on broadcast data.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	0.625	1.432	11.428
English	1.101	9.014	8.714
French	1.250	0.949	12.333
Hindi	0.654	10.878	7.666
Korean	0.148	4.563	3.238
Mandarin	0.803	7.139	5.619
Spanish	1.011	3.580	2.523
Vietnamese	0.148	3.556	9.761
Average	0.718	5.139	7.660
pooled minDET	1.104	5.447	8.166
pooled EER	1.145	5.644	8.250
pooled unweighted minDET	1.196	6.746	8.166
pooled unweighted EER	1.250	6.959	8.250

Table A.8: Results of acoustic system trained on broadcast database with channel compensation trained on broadcast data.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	15.565	17.015	0.142
English	6.517	15.538	0.476
French	11.458	13.324	0.238
Hindi	14.851	17.612	0.000
Korean	6.994	10.583	0.809
Mandarin	6.666	16.875	0.000
Spanish	6.428	10.113	0.714
Vietnamese	2.678	9.773	0.142
Average	8.895	13.854	0.315
pooled minDET	9.471	14.510	0.505
pooled EER	9.840	15.013	0.583
pooled unweighted minDET	9.196	15.939	0.505
pooled unweighted EER	9.508	16.198	0.583

Table A.9: Results of acoustic system trained on CallFriend database with channel compensation trained on telephone data.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	0.178	1.471	3.190
English	0.416	6.850	4.619
French	0.446	0.587	4.619
Hindi	0.178	7.643	2.190
Korean	0.208	2.923	1.285
Mandarin	0.505	8.168	1.285
Spanish	0.119	2.636	0.380
Vietnamese	0.000	2.005	1.142
Average	0.256	4.035	2.339
pooled minDET	0.383	4.258	2.964
pooled EER	0.420	4.296	3.083
pooled unweighted minDET	0.437	5.714	2.964
pooled unweighted EER	0.500	5.730	3.083

Table A.10: Results of acoustic system trained on broadcast database with channel compensation trained on telephone data.

Language	**NIST 2003**	**NIST 2007**	**Broadcast**
Dari	11.220	17.892	0.047
English	7.410	17.296	1.666
French	13.839	13.585	0.238
Hindi	14.970	15.280	0.095
Korean	3.720	6.250	0.714
Mandarin	9.166	20.584	0.000
Spanish	6.726	9.027	0.619
Vietnamese	2.023	5.336	0.000
Average	8.634	13.156	0.422
pooled minDET	9.136	14.136	0.886
pooled EER	9.222	14.290	0.922
pooled unweighted minDET	8.964	15.772	0.886
pooled unweighted EER	9.107	15.860	0.922

Appendix B

Cooperation with Linguistic Data Consortium

We were collaborating with the Linguistic Data Consortium (LDC) on preparation of broadcast data database, which will contain recording from various radio stations in many languages. Language labels of all recordings in this database need to be manually verified. Verification of such large amount of data consisting of tens of languages represents a problem in routing a recordings to an annotator, able to recognize language of particular recording.

We received a set of various broadcast recordings from LDC without language labels. It was expected, that these recordings contain 39 different languages. [1] This package contained over 7GB or 10150 files of stereo recordings compressed in mp3 format. Given the fact, that the recordings often contain different broadcast stations in the left and right channel, more than 14000 hours of data had to be processed and labeled.

In order to label the data, we downloaded large amount of broadcast data from the Voice of America archive, where the recordings **are labeled** according to location of broadcasting and predominant language. We prepared the data for training using the same techniques explained in section 3.1 and trained a phonotactic system based on string output from our Hungarian phoneme recognizer. The language models were trained for 43 languages [2] and there was an average of 14.1 hours of speech per each language for training. However this number varied from 4.7 hours (for Serbian)to 64 hours (for Korean).

We provided three top-scoring language labels for each file and each channel to speed up the routing of files to human annotators. We also provided speech and nonspeech labels and labels for the phone calls detected in the broadcasts. These labels were obtained by techniques explained in sections 2.1, 2.2 and 3.

We have also created software packages for phone call detection and speech/nonspeech segmentation. This software was shipped to LDC will allow them to process the recorded broadcast more effectively.

[1] Albanian, Amharic, Armenian, Azeri, Bengali, Bosnian, Burmese, Cantonese, Creole, Croatian, Dari, English, French, Georgian, Greek, Hausa, Hindi, Indonesian, Khmer, Korean, Kurdish, Lao, Mandarin, Pushto, Persian, Portuguese, Russian, Serbian, Shona, Somali, Spanish, Swahili, Thai, Tigrigna, Turkish, Ukrainian, Urdu, Uzbek, Vietnamese

[2] Albanian, Amharic, Azerbaijani, Bengali, Bosnian, Burmese, Cantonese, Creole, Croatian, Dari (Persian), English, French, Georgian, Greek, Hausa, Hindi, Indonesian, Khmer, Kinyarwanda, Korean, Kurdish, Lao, Macedonian, Mandarin, Ndebele, Oromo, Pashto, Persian, Portuguese, Russian, Serbian, Shona, Somali, Spanish, Swahili, Thai, Tibetan, Tigrinya, "Talk To America - English", Turkish, Ukrainian, Urdu, Uzbek, Vietnamese

Bibliography

[1] Ivo Řezníček, "Audiovisual recording system," Diploma thesis, Brno University of Technology FIT, 2007.

[2] Petr Schwarz, Pavel Matějka, and Jan Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proceedings of ICASSP 2006*, 2006, pp. 325–328.

[3] Petr Schwarz, Pavel Matějka, and Jan Černocký, "Towards lower error rates in phoneme recognition," in *Proceedings of 7th International Conference Text,Speech and Dialoque*, 2004.

[4] "The 2003 NIST Language Recognition Evaluation Plan (LRE03)," http://www.nist.gov/speech/tests/lre/2003/LRE03EvalPlan-v1.pdf.

[5] "The 2007 NIST Language Recognition Evaluation Plan (LRE07)," http://www.nist.gov/speech/tests/lang/2007/LRE07EvalPlan-v8b.pdf.

[6] Pavel Matějka, Lukáš Burget, Ondřej Glembek, Petr Schwarz, Valiantsina Hubeika, Michal Fapšo, Tomáš Mikolov, and Oldřich Plchot, "But system description for nist lre 2007," in *Proc. 2007 NIST Language Recognition Evaluation Workshop*. 2007, pp. 1–5, National Institute of Standards and Technology.

[7] Ondej Glembek, Pavel Matjka, Luk Burget, and Tom Mikolov, "Advances in phonotactic language recognition," in *Proc. Interspeech 2008*. 2008, p. 4, International Speech Communication Association.

[8] Niko Brummer and David van Leeuwen, "On calibration of language recognition scores," in *Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, June 2006, pp. 1–8.

[9] Lukáš Burget, Pavel Matějka, Petr Schwarz, Ondřej Glembek, and Jan Černocký, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1979–1986, 2007.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, jan 2000.

[11] P. Matějka, L. Burget, P. Schwarz, and J. Černocký, "Brno university of technology system for nist 2005 language recognition evaluation," in *Proc. NIST LRE 2005 Workshop*, San Juan, Puerto Rico, June 2006, pp. 57–64.

[12] Pedro A. Torres-Carrasquillo, Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds, J.R. Deller, and Jr., "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. 7 th International Conference on Spoken Language Processing*, Denver, Colorado, USA, Sept. 2002.

[13] Jordan Cohen, Terri Kamm, and Andreas G. Andreou, "Vocal tract normalization in speech recognition: Compensating for systematic speaker variability," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3246–3247, 1995.

[14] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech*, Rhodes, Greece, Sept. 1997, pp. 963–966.

[15] Niko Brummer, "Spescom DataVoice NIST 2004 system description," in *Proc. NIST Speaker Recognition Evaluation 2004*, Toledo, Spain, June 2004.