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

Oldřich Plchot, Valiantsina Hubeika, Lukáš Burget, Petr Schwarz and Pavel Matějka

Speech@FIT, Brno University of Technology, Czech Republic,
iplchot|burget|schwarzp|matejkap@fit.vutbr.cz, xhubei00@stud.fit.vutbr.cz

Abstract. This paper presents a procedure of acquiring linguistic data from the broadcast media and its use in language recognition. The goal of this work is to answer the question whether the automatically obtained data from broadcasts can replace or augment to the continuous telephone speech. The main challenges are channel compensation issues and great portion of unspontaneous speech in broadcasts. The experimental results are obtained on NIST LRE 2007 evaluation system, using both NIST provided training data and data, obtained from broadcasts.

Key words: Language Identification (LID), Broadcast data, Phone call detection, Channel compensation

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 exists.

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 should provide us with large

[*] This work was partly supported by European projects AMIDA (FP6-033812), Caretaker (FP6-027231) and MOBIO (FP7-214324), by Grant Agency of Czech Republic under project No. 102/08/0707 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under project No. 162/2005. Lukáš Burget was supported by Grant Agency of Czech Republic under project No. GP102/06/383.

Table 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

amount of data that we expect will lead to improved performance for languages included in our present systems and to capability of processing languages that we were unable to recognize due to absence of the 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 both sets of obtained data (wideband speech segments and phone calls) by training and evaluating the systems for languages with a lack of standard training data and on the basis of the results conclude, whether these data can be used to improve existing LRE systems.

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 with Thai language so far and we are planning to extend this work to the broad range of other languages.

2 Data Acquisition

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 the Czech Republic are shown in Table 1 [1].

All of the listed sources except Internet radios are geographically dependent regarding location. The quality of different Internet sources vary a lot and it is important to carefully choose them. We have used an archive[1] of Voice of America Internet radio to obtain the Thai data, as the Thai language is not available from DVB-S in the Czech Republic.

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

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

we obtained 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 constantly. However we have not investigated this potential problem so far.

3 Experiment

3.1 Selection of Scenario

Two main scenarios face the issue of the lack of training data. It can be either the situation when we have not enough training data from standard sources or when we do not have any data for particular language at all. In NIST LRE 2007 evaluation we can simulate the first scenario with Thai where we have only 1.5 hour of training data (see Table 2), and we can easily simulate the second scenario by removing these data.

We downloaded about 250 hours of radio data. We decided to select only phone calls, because they match the target CTS data and we can successfully detect them. The phone calls usually contain no music and we suppose they represent conversational speech. We detected 10 hours of telephone calls in the obtained data which is about 4% of the original recordings. Since there were some English calls, the phonotactic LID (from LRE2005 BUT system [2]) was used to detect Thai versus English. Consequently we listened to all samples with low recognition confidence to verify they are not English.

3.2 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

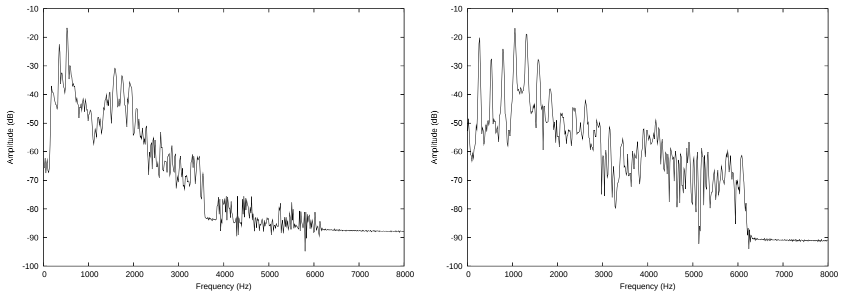

Fig. 1. Power Spectral Density of telephone call in the broadcast (left figure) and wideband speech (right figure).

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 (see Figure 2).

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 1). At first the PSD was normalized to zero mean and unit variance. Then values in the first half and values in the second half of the PSD were summed. A ratio between these two sums was compared with a threshold and the decision was made. If the sum from higher frequencies is bigger than the sum from lower frequencies, there is more energy and it is a wide band speech.

3.3 System Description

We use the best single performing system from our NIST LRE 2007 submission to perform the following experiments. The inspiration for this system came from our GMM system for speaker recognition [3] which follows conventional Universal Background Model-Gaussian Mixture Modeling (UBM-GMM) paradigm [4] and employs number of techniques that have previously proved to improve GMM system performance [5]. This system was chosen because it can easily compensate for the channel distortion.

Table 2 lists the corpora (distributed by LDC and ELRA) used to train our systems.

Our system uses the popular shifted-delta-cepstra (SDC) [6] 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) [7] 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

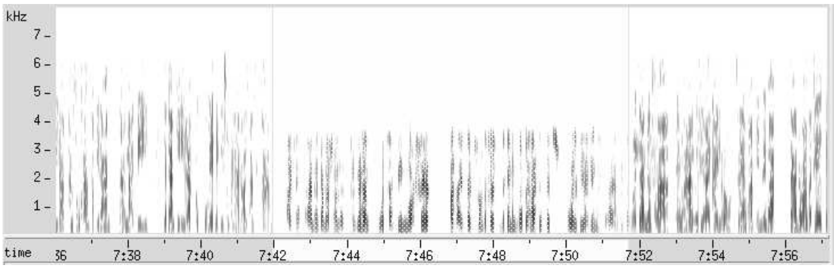

Fig. 2. Phone Call in a Radio Broadcast.

Table 2. Training data in hours for each language and source.

	sum	CF	CH	F	SRE	LDC07	OGI	OGI22	Other
Arabic	212	19.5	10.4	175	5.93	1.45		0.33	
Bengali	4.27				2.86	1.42			
Chinese	93.2	41.7	1.64	17.2	44.9	4.2	0.87	0.85	
English	264	39.8	4.68	162	34.9		6.77	0.52	15.6 (FAE)
Hindustani	23.5	19.6			0.64	1.32	1.53	0.42	
Spanish	54.3	43.8	6.71		2.63		1.18	0.38	
Farsi	22.7	21.2			0.03		1.00	0.42	
German	28.2	21.6	5.10				1.12	0.38	
Japanese	23.9	19.1	3.47				0.87	0.35	
Korean	19.7	18.4			0.09		0.72	0.5	
Russian	15.1				3.38	1.33		0.43	10.0 (SpDat)
Tamil	19.6	18.4					0.96	0.26	
Thai	**1.45**				0.15	1.23			
Vietnamese	21.6	20.6					0.79	0.27	
Other	62.5	20.7					1.10	3.29	37.4 (SpDat)

CF	CallFriend
CH	CallHome
F	Fisher English Part 1.and 2.
F	Fisher Levantine Arabic
F	HKUST Mandarin
SRE	Mixer (data from NIST SRE 2004, 2005, 2006)
LDC07	development data for NIST LRE 2007
OGI	OGI-multilingual
OGI22	OGI 22 languages
FAE	Foreigen Accented English
SpDat	SpeechDat-East
SB	SwitchBoard

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 [8] of UBM using enrollment conversation. Only means are adapted.

In verification phase, standard Top-N Expected Log Likelihood Ratio (ELLR) scoring [8] 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 [3] prior to computing the log likelihood ratio score.

3.4 Channel compensation

As the data used in the system were obtained from different databases and therefore recorded over different channels, eigenchannel adaptation [9] was applied in order to compensate the channel distortion. In language detection task, channel

Table 3. Results. Channel comp. stands for the system with eigenchannel compensation. Channel comp. +4 stands for the system where to the former eigenchannels 4 additional eigenchannels computed only on Thai data (NIST and Radio) were added.

	No channel comp.		Channel comp.		Channel comp. +4
	NIST	Radio	NIST	Radio	Radio
DCF all lang.	12.83	13.66	7.30	7.56	7.47
EER all lang.	13.02	13.80	7.41	7.65	7.59
Thai DCF	7.81	11.61	3.93	6.05	5.97

variability may comprehend not only variability in the telephone channel or type of microphone but also session or speaker variability.

4 Results and discussion

The results are evaluated using standard metrics: Detection Error Tradeoff (DET) curve (see Figure 3), Decision Cost Function (DCF) and Equal Error Rate (EER) [10]. Several experiments were run using the original telephone data provided by NIST and the telephone data acquired from radio. All experiments were done on *10 second segments*. Two Thai models were trained on both sets respectively. Other 13 language-dependent GMMs were shared by both systems.The results are presented in Table 3 for the Thai language only (Thai DCF row) and for the complete systems containing all 14languages. The “NIST” column stands for a system with Thai trained on 1.45 hours of CTS data available from LDC, the “Radio”denotes systems trained on 10 hours telephone data obtained from broadcasts.

First, both systems were trained without channel compensation. Then, eigenchannel adaptation was applied, using a matrix containing 50 eigenchannels computed without radio data.

The third experiment was performed by adding several eigenchannels estimated using only Thai data (both, CTS and radio data) and concatenated to the former eigenchannel matrix. This approach decreased the error in the system where Thai was trained on radio data, although this decrease was not significant.

However the recognizer trained on radio data does not bring as good results as the recognizer trained on CTS data, the results show (in comparison to the average DCF) that in case of language there are no available data for, radio data can be used.

5 Conclusions and future work

We introduced a simple but promising approach of acquiring telephone data for LID. Experiments with Thai language modeled using standard telephone data and telephone data acquired from broadcast were performed. Although the

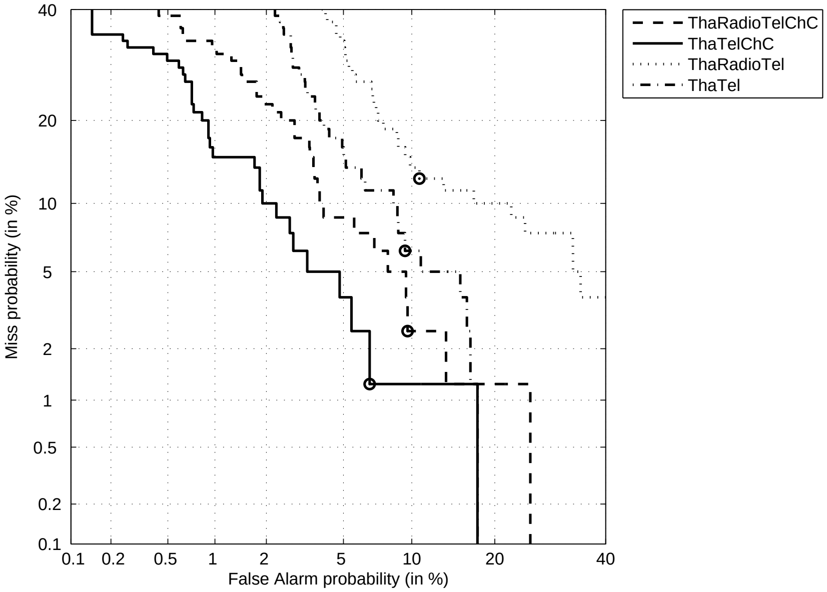

Fig. 3. Results achieved with both systems with and without channel compensation for Thai only. The det curves marked as ThaiTel and ThaiTelChC represent the results achieved on the original telephone data (CTS) with and without eigenchannel adaptation respectively. The det curves marked as ThaiRadioTel and ThaiRadioTelChC represent the results achieved on the telephone data acquired from radio with and without eigenchannel adaptation respectively.

system trained on this data did not outperform the one trained on CTS, the degradation was not critical and we consider it a viable option for scenarios with very little or no data for a given language. Future investigation on more languages should be performed. Our main goal will be to obtain the radio data for all 14 languages and to train LID system on these data. We will also investigate into channel compensation techniques in cases several sources (CTS, broadcast-telephone, read speech) are mixed.

References

1. Ivo Řezníček, "Audiovisual recording system," Diploma thesis, Brno University of Technology FIT, 2007.
2. Pavel Matějka, Lukáš Burget, Petr Schwarz, and Jan Černocký, "Nist language recognition evaluation 2005," in *Proceedings of NIST LRE 2005*, 2006, pp. 1–37.
3. 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.

4. 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.
5. 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.
6. 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.
7. 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.
8. D. A. Reynolds, “Comparison of background normalization methods for text-independent speaker verification,” in *Proc. Eurospeech*, Rhodes, Greece, Sept. 1997, pp. 963–966.
9. Niko Brummer, “Spescom DataVoice NIST 2004 system description,” in *Proc. NIST Speaker Recognition Evaluation 2004*, Toledo, Spain, June 2004.
10. “The 2007 NIST Language Recognition Evaluation Plan (LRE07),” http://www.nist.gov/speech/tests/lang/2007/LRE07EvalPlan-v8b.pdf.