IDIAP RESEARCH REPORT

MEDIAPARL: BILINGUAL MIXED LANGUAGE ACCENTED SPEECH DATABASE

David Imseng Hervé Bourlard Holger Caesar
Philip N. Garner Gwénolé Lecorvé
Alexandre Nanchen

Idiap-RR-03-2013

JANUARY 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

MediaParl: Bilingual mixed language accented speech database

David Imseng, Hervé Bourlard, Holger Caesar,
Philip N. Garner, Gwénolé Lecorvé, Alexandre Nanchen

January 4, 2013

Abstract

MediaParl is a Swiss accented bilingual database containing recordings in both French and German as they are spoken in Switzerland. The data were recorded at the Valais Parliament. Valais is a bilingual Swiss canton with many local accents and dialects. Therefore, the database contains data with high variability and is suitable to study multilingual, accented and non-native speech recognition as well as language identification and language switch detection.

We also define monolingual and mixed language automatic speech recognition and language identifictaion tasks and evaluate baseline systems.

The database is publicly available for download.

1 Introduction

In this paper, we present a database that addresses multilingual, accented and non-native speech, which are still challenging tasks for current ASR systems. At least two bilingual databases already exist (Lyu et al., 2010; Alabau and Martinez, 2006). The MediaParl speech corpus was recorded in Valais, a bilingual canton of Switzerland. Valais is surrounded by mountains and is better known internationally for its ski resorts like Verbier and Zermatt with the Matterhorn. Valais is an ideal place to record bilingual data, because there are two different official languages (French and German). Furthermore, even within Valais, there are many local accents and dialects (especially in the German speaking part). This language mix leads to obvious difficulties, with many people working and even living in a non-native language, and leads to high variability in the speech recordings. On the other hand, it leads to valuable data that allow study of multilingual, accented and non-native speech as well as language identification and language switch detection.

MediaParl was recorded at the cantonal parliament of Valais. About two thirds of the population speak French, and one third speaks German. However, the German that is spoken in Valais is a group of dialects (also known as *Walliser Deutsch*) without written form. The dialects differ a lot from the standard (high) German (Hochdeutsch, spoken in Germany) and are sometimes even difficult to understand for other Swiss Germans. Close to the language border (Italy and French speaking Valais) people also use foreign words (loan words) in their dialect. In the parliament (and other formal situations), people speak in accented standard German. In the remainder of the paper, we refer simply to German and French but take this to mean Swiss German and Swiss French.

The political debates at the parliament are recorded and broadcasted. The recordings mostly contain prepared speeches in both languages. Some of the speakers even switch between the two languages during the speech. Therefore, the database may also be used to a certain extent to study code-switched ASR. However, in contrast to for example (Lyu et al., 2010), the code switches always occur on sentence boundaries. While some similar databases only contain one hour of speech per language (Alabau and Martinez, 2006), MediaParl contains 20 hours of German and 20 hours of French data.

In the remainder of the paper, we will give more details about the recording (Section 2) and transcription (Section 3) process. We will also present the dictionary creation process in Section 4 and define tasks and training, development and test sets in Section 5. Finally we present and evaluate baseline systems on most of the tasks in Section 6.

2 Recordings

The MediaParl speech corpus was recorded at the cantonal parliament of Valais, Switzerland. We used the recordings of Swiss Valaisan parliament debates of the years 2006 and 2009. The parliament debates always take place in the same closed room. Each speaker intervention can last from about 10 seconds up to 15 minutes. Speakers are sitting or standing when talking and their voice is recorded through a distant microphone. The recordings from 2009 that were processed at Idiap Research Institute are also available as video streams online[1].

The audio recordings of the year 2006 were formatted as "mp3", more specifically MPEG ADTS, layer III, v1, 128 kbps, 44.1 kHz, Monaural with16 bits per sample. The video recordings of the year 2009 were formatted as "avi" with uncompressed PCM (stereo, 48000 Hz, 16 bits per sample) audio data. All the audio data (2006 and 2009) was converted to WAVE audio, Microsoft PCM, 16 bit, mono 16000 Hz prior to any processing.

3 Transcriptions

Each recorded political debate (session) lasts about 3 hours and human-generated transcriptions are available. However, manual work was required to obtain annotated speech data of reasonable quality:

- Speaker diarization was performed manually. Each speaker cluster is referred to as an *intervention*; an intervention consists of multiple sentences consecutively spoken by the same speaker.
- Each intervention is then associated with the corresponding transcription and manually split into individual sentences.
- The transcription of each sentence is then manually verified by two annotators and noisy utterances are discarded.

Finally, all the transcriptions were tokenized and normalized using in-house scripts. The transcription process described above resulted in a corpus of 7,042 annotated sentences (about 20 hours of speech) for the French language and 8,526 sentences (also about 20 hours of speech) for the German language.

4 Dictionaries

The phonemes in the dictionaries are represented using the Speech Assessment Methods Phonetic Alphabet (SAMPA)[2]. SAMPA is based on the International Phonetic Alphabet (IPA), but features only ASCII characters. It supports multiple languages including German and French.

Manual creation of a dictionary can be quite time consuming because it requires a language expert to expand each word into its pronunciation. Therefore we bootstrap our dictionaries with publicly available sources that are designed for a general domain of speech, such as conversations. However, the speech corpus that we use includes large numbers of words, that are specific to the domain (politics) and region (Switzerland). Hence, the dictionaries need to be completed. In the reminder of this section, we first generally describe how we complete the dictionaries and then give more details about German and French in Sections 4.2 and 4.3 respectively.

4.1 Phonetisaurus

We used Phonetisaurus (Novak et al., 2012), a grapheme-to-phoneme (g2p) tool that uses existing dictionaries to derive a finite state transducer based mapping of sequences of letters (graphemes) to their acoustic representation (phonemes). The transducer was then applied to unseen words.

For languages with highly transparent orthographies such as Spanish or German (Goswani, 1999), g2p approaches typically work quite well (Schlippe et al., 2012). However, for languages with less transparent orthographies, such as English or French (Goswani, 1999), it is relatively difficult to derive simple mappings from the grapheme representation of a syllable to its phoneme representation. Therefore, g2p approaches tend to work less well (Schlippe et al., 2012).

[1] http://www.canal9.ch/television-valaisanne/emissions/grand-conseil.html
[2] http://www.phon.ucl.ac.uk/home/sampa/index.html

Furthermore, due to the prevalence of English in many fields, domain-specific words, such as "high-speed", "interview" or "controlling" are often borrowed from English. Since MediaParl was recorded in a bilingual region, this effect becomes even more pronounced than in other more homogenous speaker populations. As a result, the dictionaries contain relatively large numbers of foreign words. However, the g2p mappings of one language do not necessarily generalize to a foreign language. French word suffixes for example, are often not pronounced if they form an extension to the word stem, such as plurals and conjugations. On the other hand, German word suffixes are usually pronounced, except for some cases where terminal devoicing (voiced consonants become unvoiced before vowels or breaks) applies.

Owing to the above problems with g2p, all entries generated by Phonetisaurus were manually verified by native speakers according to the SAMPA rules for the respective language. Table 2 shows the number of unique words in each dictionary.

4.2 German Dictionary

To bootstrap the German dictionary, we used Phonolex[3]. Phonolex was developed by a cooperation between DFKI Saarbrücken, the Computational Linguistics Lab, the Universität Leipzig (UL) and the Bavarian Archive for Speech Signals (BAS) in Munich.

82% of the German MediaParl words were found in Phonolex. Phonetisaurus was then trained on Phonolex to generate the missing pronunciations. All g2p-based dictionary entries were manually verified in accordance to the German SAMPA rules (Caesar, 2012).

Since Phonolex is a standard German dictionary and we only use one pronunciation for each word, the actual Swiss German pronunciation of some words may significantly differ. Analyzing, for instance, various samples of the German word "achtzig" reveals that speakers in MediaParl pronounce it in three different ways:

1. /Q a x t s I C/
2. /Q a x t s I k/
3. /Q a x t s I k C/

where (1) is the standard German version used in Phonolex, (2) can be found in various German dialects and (3) seems to be a Swiss German peculiarity.

4.3 French Dictionary

The French dictionary was bootstrapped with BDLEX[4]. BDLEX consists of a lexical database developed at Institut de Recherche en Informatique (IRIT) in Toulouse. The data cover lexical, phonological, and morphological information.

83% of the French MediaParl words were found in BDLEX. Similar to German, we trained Phonetisaurus on BDLEX to generate the missing pronunciations. Again, all g2p-based dictionary entries were manually verified in accordance to the French SAMPA rules.

5 Definitions

In this section, we first define tasks that can be performed on the database and then present the partition of the database into training, development and test data.

5.1 Tasks

The database is well suited to study the following tasks:

Automatic speech recognition (ASR) The ASR task consists of performing monolingual independent ASR for French and German. As usually done, the performance can be measured with word accuracies.

The database is particularly well suited to investigate non-native ASR and we will see in Section 6 that ASR on non-native utterances is more challenging.

[3] http://www.phonetik.uni-muenchen.de/Bas/BasPHONOLEXeng.html
[4] http://catalog.elra.info/product_info.php?products_id=33

Speaker	Sentences in French	Sentences in German
059	31	195
079	22	698
094	313	72
096	89	8
102	72	7
109	233	402
191	165	310
Total	925	1692

Table 1: MediaParl-TST: speakers using both languages form the test set. For each speaker the number of French and German sentences is given.

Language identification (LID) The LID task consists of determining the spoken language for each sentence. In that case, the performance can be measured simply as percentage of sentences for which the spoken language was correctly recognized because the decision is either correct or wrong.

As already described in Section 3, an intervention contains multiple sentences of the same speaker. The bilingual speakers change language within one intervention, hence the database can also be used to study the detection of language switches. Note that the language switches always occur at sentence boundaries.

Mixed language ASR Mixed language ASR is defined as ASR without knowing the language of a sentence a priori. As for the ASR task, performance can be measured in word accuracies. The mixed language ASR task is considered to be much more challenging than the standard ASR task.

Since interventions contain language switches, the database may also be used to investigate code-switched ASR. However, note that the language switches always happen at sentence borders what is simpler than code-switched ASR as defined by for example (Lyu et al., 2010).

Speaker diarization The whole database is labeled with speaker information. Therefore it may also be used to perform speaker diarization. Furthermore, many speakers can be found in multiple interventions, hence speaker diarization might also be applied across interventions.

5.2 Data partitioning

We partitioned the database into training, development and test sets. Since we focus on bilingual (accented, non-native) speech, the test set (MediaParl-TST) contains all the speakers which speak in both languages (see Table 1). Hence, MediaParl-TST contains all the non-native utterances. 90% of the remaining speakers (only speaking in one language) form the training set (MediaParl-TRN) and the other 10% the development set (MediaParl-DEV). Training and development speakers were randomly determined.

MediaParl-TRN contains 11,425 sentences (5,471 in French and 5,955 in German) spoken by 180 different speakers and MediaParl-DEV contains 1,525 sentences (646 in French and 879 in German) from 17 different speakers. The speakers from MediaParl-TST are shown in Table 1. As already described, each speaker uses both languages. Table 1 also displays how many French and German sentences were recorded for each test speaker. We assume that each speaker is naturally speaking more often in his mother tongue. Hence, the speakers 059, 079, 109 and 191 appear to be native German speakers and the speakers 094, 096 and 102 native French speakers. These findings were confirmed by native speakers of French and German. The speakers 109 and 191 are native German speakers but they are very fluent in the second language.

6 Baseline systems

In this section, we present baseline systems for some of the aforementioned tasks. First, we describe the acoustic feature extraction process and then present ASR, LID and mixed ASR results.

Language	Vocabulary size	Number of bigrams	Perplexities	
			DEV	TST
French	12,035	1.5 M	147	152
German	16,727	1.9 M	295	360

Table 2: Statistics of the monolingual language models.

Speaker	French	+/- avg.	German	+/- avg.
059	39.7%	-43.7%	70.8%	+3.5%
079	42.7%	-39.4%	68.7%	+0.4%
109	69.2%	-1.8%	74.2%	+8.5%
191	54.1%	-23.3%	60.1%	-12.1%
094	78.7%	+11.6%	66.8%	-2.3%
096	79.2%	+12.3%	63.1%	-7.7%
102	78.3%	+11.1%	51.4%	-24.9%
Avg	70.5 %	—	68.4%	—

Table 3: ASR performance of the different speakers. The relative change compared to the average performance is also given. Speakers 059, 079, 109 and 191 are considered as native German speakers and the others as native French speakers.

6.1 Feature extraction

For all the experiments presented in this paper, we used 39 Mel-Frequency Perceptual Linear Prediction (MF-PLP) features (C0-C12+Δ+$\Delta\Delta$), extracted with the HTS variant[5] of the HTK toolkit.

6.2 Automatic Speech Recognition

For the ASR task, we built two independent ASR systems, one for French and one for German. Context dependent triphone Gaussian mixture models were trained using HTS. Each triphone was modeled with three states and each state was modeled with 16 Gaussians. To tie rare states, we applied a conventional decision tree. The minimum description length criterion was used to determine the number of tied states (Shinoda and Watanabe, 1997).

Bigram language models were independently trained for French and German. For each language, two sources were considered for bigram probabilities estimation: the transcriptions of the training set and texts from the corpus Europarl, a multilingual corpus of European Parliament proceedings (Koehn, 2005). Europarl is made up of about 50 million words for each language and is used to overcome data sparsity of the MediaParl texts. However, vocabularies were limited to the sole words from MediaParl, including words from the development and test sets in order to avoid out-of-vocabulary word problems in the experiments. Statistics from both sources were smoothed using Witten-Bell smoothing and were then linearly interpolated. Interpolation weights were tuned by minimizing the perplexity of the transcriptions of the development set and no pruning was applied. Sizes and perplexities of these monolingual language models are summarized in Table 2.

Decoding was performed with HTS. Language model scaling factor and insertion penalty were tuned on the development sets.

We hypothesized that the standard speech recognition systems would perform worse on non-native speech, owing to the accent associated with even fluent speakers. Table 3 shows the performance of the HMM/GMM systems on French and German respectively for each speaker. It can clearly be seen that the German native speakers perform worse on the French data and vice versa, hence our hypothesis is confirmed. The effect seems to be more pronounced on the French data. This might have to do with the German dictionary, which seems to be suboptimal since it is based on standard German and not on the dialect.

[5] http://hts.sp.nitech.ac.jp/

Speaker	French Data	German Data	All Data
059	83.9%	100%	97.8%
079	90.9%	99.4%	99.2%
109	98.7%	100%	99.5%
191	93.4%	99.4%	97.5%
094	96.8%	100%	97.4%
096	100%	87.5%	99.0%
102	97.2%	85.7%	96.2%
Avg	96.5%	99.5%	98.5%

Table 4: LID performance for the different speakers. The performance on all the test data is given in the rightmost column. The results are also split into French and German data. The system performs better on the native speech for all speakers except 094.

6.3 Language identification

To perform LID, we applied the recently proposed hierarchical multilayer perceptron (MLP) based language identification approach (Imseng et al., 2010). The first layer of the hierarchical MLP classifier is a shared phoneme set MLP classifier which was trained on French and German data. The resulting (bilingual) posterior sequence is fed into a second MLP taking a larger temporal context into account. The second MLP can learn implicitly different types of patterns such as confusion between phonemes and phonotactics for LID.

To train the shared phoneme set MLP classifier, we built a shared phoneme set by merging French and German phonemes if they were represented by the same SAMPA symbol. We used nine frames temporal context (one frame every 10 ms) as input for the MLP. Following a common strategy, the number of hidden units was determined by fixing the number of parameters to 10% of the total number of training samples. As already mentioned, the second MLP was then trained on a larger temporal context. In this study, we used 29 frames. The outputs of the second MLP are language posterior probabilities given the acoustics at the input. Given a test utterance, the frame-based log posteriors for each language are summed up and a decision about the language is made by choosing the language that gets the maximum log posterior probability over the whole utterance.

We hypothesized that the LID performance on non-native speech would be lower, for much the same reasons as for ASR. The results of the language identification system can be found in Table 4. The results are split into French and German data. The LID performance is always better on data of the speaker's mother tongue except for speaker 094, who is a native French speaker. Hence our hypothesis is confirmed. The lower overall performance on French data may be explained by the fact that 49% of the sentences are non-native speech, whereas only 5% of the German sentences are non-native speech.

6.4 Mixed language ASR

To perform mixed language ASR, we used two different approaches:

Shared system We built one multilingual decoder trained on the data of both languages. To build the shared system, we first created a shared phoneme set that contains all the German and French phonemes. As we did for the hierarchical LID approach, we merged phonemes that share the same SAMPA symbol. Then we trained GMMs as described for the monolingual systems in Section 6.2.

Our multilingual language modeling is similar to an approach presented in (Wang et al., 2002). More specifically, all words of training texts used in Section 6.2 and entries of the French and German vocabularies were first labeled with tags corresponding to their respective language[6]. Then, the multilingual vocabulary is defined as the union of tagged monolingual vocabularies. Finally, monolingual bigram probabilities were trained on the tagged texts and linearly interpolated such that each language shared the same probability mass.

Perplexities of the multilingual language model on the French and German parts of the development and test sets are presented in Table 5. This preliminary approach is not optimal since the sizes of the vocabularies are not exactly the same. Therefore, in our experiments, the probability of a German word is on average lower than that of a French word. Incorporating this mismatch within linear interpolation should provide better performance.

[6]For instance, French words are suffixed with the string `_fr`, and German ones with `_de`.

	DEV	TST
French	279	289
German	554	661

Table 5: Perplexities of the multilingual language model on the French and German parts of MediaParl-DEV and MediaParl-TST.

Speaker	Shared System	Language Switch	Oracle LID
059	60.3%	66.1%	66.9%
079	59.3%	67.7%	67.9%
109	65.6%	71.9%	72.0%
191	50.4%	57.2%	57.8%
094	70.5%	76.8%	77.4%
096	74.4%	78.5%	78.5%
102	73.2%	77.0%	77.5%
Avg	62.5%	69.0%	69.4%

Table 6: ASR performance of the different speakers. The performance of the shared system, the language switch system and a language switch with oracle LID is given.

Language switch For this system we first performed LID as described in Section 6.3 and then used the respective monolingual decoder from Section 6.2. For the sake of comparison, we also evaluated a system with oracle LID, i.e., a system where we know the language in advance and pick the correct monolingual recognizer.

Obviously, the oracle LID system will perform better than the language switch system because the LID errors cannot be corrected after the wrong decoder is chosen. We hypothesized that the language switch system would outperform the shared system because we have already seen in Section 6.3 that the LID performance is close to 100%.

Table 6 confirms our hypothesis and shows the mixed language ASR performance for each speaker.

7 Public distribution

We have presented a bilingual mixed language accented database that contains French and German data recorded at the Valais Parliament. The test set contains all the speakers that use both languages during the political debates. We also presented baseline systems for ASR, LID and mixed language ASR.

We are happy to announce that this database is publicly available through http://www.idiap.ch/dataset/mediaparl. The database contains the raw audio recordings, the transcriptions (word level) and the file lists for MediaParl-TRN, MediaParl-DEV and MediaParl-TST. The dictionaries are derived from BDLex and Phonolex, and hence cannot be provided directly. However, they can be generated automatically using scripts provided if those base dictionaries are available and the grapheme-to-phoneme tool is installed. The dictionaries are distributed through ELRA as ELRA-S0004 and ELRA-S0035 respectively; the required software tool, phonetisaurus, is available online[7].

8 Acknowledgement

We are grateful to the Parliament Service of the State of Valais for providing access to the parliament debate A/V recordings.

References

V. Alabau and C. Martinez. Bilingual speech corpus in two phonetically similar languages. In *International Conference on Language Resources and Evaluation*, pages 1624–1627, 2006.

[7] http://code.google.com/p/phonetisaurus/

Holger Caesar. Integrating language identification to improve multilingual speech recognition. Technical Report Idiap-RR-24-2012, Idiap Research Institute, July 2012.

Usha Goswani. *The relationship between phonological awareness and orthographic representation in different orthographies*, chapter 8. Cambridge University Press, 1999.

D. Imseng, M. Magimai.-Doss, and H. Bourlard. Hierarchical multilayer perceptron based language identification. In *Proc. of Interspeech*, pages 2722–2725, 2010.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proc. of the 10th Machine Translation Summit*, pages 79–86, 2005.

Dau-Cheng Lyu et al. SEAME: a Mandarin-English code-switching speech corpus in south-east asia. In *Proc. of Interspeech*, pages 1986–1989, 2010.

J. Novak et al. Improving WFST-based G2P conversion with alignment constraints and RNNLM N-best rescoring. In *Proc. of Interspeech*, 2012.

T. Schlippe, S Ochs, and T. Schultz. Grapheme-to-phoneme model generation for indo-european languages. In *Proc. of ICASSP*, pages 4801–4804, 2012.

Koichi Shinoda and Takao Watanabe. Acoustic modeling based on the MDL principle for speech recognition. In *Proc. of Eurospeech*, volume I, pages 99–102, 1997.

Z. Wang, U. Topkara, T. Schultz, and A. Waibel. Towards universal speech recognition. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, pages 247–252, 2002.