GlobalPhone: A Multilingual Text & Speech Database in 20 Languages

Tanja Schultz, Ngoc Thang Vu, Tim Schlippe

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany

tanja.schultz@kit.edu

Abstract

This paper describes the advances in the multilingual text and speech database GlobalPhone, a multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries in 20 languages. GlobalPhone was designed to be uniform across languages with respect to the amount of data, speech quality, the collection scenario, the transcription and phone set conventions. With more than 400 hours of transcribed audio data from more than 2000 native speakers GlobalPhone supplies an excellent basis for research in the areas of multilingual speech recognition, rapid deployment of speech processing systems to yet unsupported languages, language identification tasks, speaker recognition in multiple languages, multilingual speech synthesis, as well as monolingual speech recognition in a large variety of languages.

Index Terms: Speech, Text, and Dictionary Resources for Multilingual Speech Processing

1. Introduction

With more than 6900 languages in the world [1] and the need to support multiple input and output languages, it is one of the most pressing challenge for the speech and language community to develop and deploy speech processing systems in yet unsupported languages rapidly and at reasonable costs [2, 3]. Major bottlenecks are the sparseness of speech and text data with corresponding pronunciation dictionaries, the lack of language conventions, and the gap between technology and language expertise. Data sparseness is a critical issue due to the fact that today's speech technologies heavily rely on statistically based modeling schemes, such as Hidden Markov Models and n-gram language modeling. Although statistical modeling algorithms are mostly language independent and proved to work well for a variety of languages, reliable parameter estimation requires vast amounts of training data. Large-scale data resources for research are available for less than 100 languages and the costs for these collections are prohibitive to all but the most widely spoken and economically viable languages. The lack of language conventions concerns a surprisingly large number of languages and dialects. The lack of a standardized writing system for example hinders web harvesting of large text corpora and the construction of pronunciation dictionaries and lexicons. Last but not least, despite the well-defined process of system building, it is cost- and time consuming to handle language-specific peculiarities, and requires substantial language expertise. Unfortunately, it is extremely difficult to find system developers who have both, the necessary technical background and the native expertise of a language in question. As a result, one of the pivotal issues for developing speech processing systems in multiple languages is the challenge of bridging the gap between language and technology expertise [2].

To date, the standard way of building speech applications for an unsupported language is to collect a sizable training corpus and to train statistical models for the new language from scratch. Considering the enormous number of languages and dialects in the world, this is clearly a suboptimal strategy, which highlights the need for more sophisticated modeling techniques. It would be desirable to develop models that can take advantage of similarities between dialects and languages of similar type and models that can share data across different varieties. This would have two benefits, first leading to truly multilingual speech processing which can handle common phenomenon such as code switching, and second providing models that are likely to be more robust toward dialectal and cross-lingual accent variations. These multilingual shared models can then be used as seed models to jump-start a system in an unsupported language by efficiently adapting the seeds using limited data from the language in questions. We refer to this development strategy as rapid language adaptation.

Ten years ago we released a multilingual text and speech corpus GlobalPhone to address the lack of databases which are consistent across languages [4]. By that time the database consisted of 15 languages but since then has been significantly extended to cover more languages, more speakers, more word tokens along with their pronunciations, and more text resources. In addition, GlobalPhone was adopted as a benchmark database for research and development of multilingual speech processing systems. Therefore, we believe the time is right to present the latest status of GlobalPhone. This paper summarizes the resources and systems available in 20 languages, and describes speech recognition performances to provide a reference and benchmark for researchers and developers working with this database.

2. The GlobalPhone Corpus

GlobalPhone is a multilingual data corpus developed in collaboration with the Karlsruhe Institute of Technology (KIT). The complete data corpus comprises (1) audio/speech data, i.e. high-quality recordings of spoken utterances read by native speakers, (2) corresponding transcriptions, (3) pronunciation dictionaries covering the vocabulary of the transcripts, and (4) baseline n-gram language models. The first two are referred to as GlobalPhone Speech and Text Database (GP-ST), the third as GlobalPhone Dictionaries (GP-Dict), and the latter as GlobalPhone Language Models (GP-LM). GP-ST is distributed under a research or commercial license by two authorized distributors, the European Language Resources Association (ELRA) [5] and Appen Butler Hill Pty Ltd. [6]. GP-Dict is distributed by ELRA, while the GP-LMs are freely available for download from our website [7].

The entire GlobalPhone corpus provides a multilingual database of word-level transcribed high-quality speech for the development and evaluation of large vocabulary speech processing systems in the most widespread languages of the world. GlobalPhone is designed to be uniform across languages with respect to the amount of data per language, the audio quality (microphone, noise, channel), the collection scenario (task, setup, speaking style), as well as the transcription and phone set conventions (IPA-based naming of phones in all pronunciation dictionaries). Thus, GlobalPhone supplies an excellent basis for research in the areas of (1) multilingual speech recognition, (2) rapid deployment of speech processing systems to yet unsupported languages, (3) language identification tasks, (4) speaker recognition in multiple languages, (5) multilingual speech synthesis, as well as (6) monolingual speech recognition in a large variety of languages.

2.1. Language Coverage

To date, the GlobalPhone corpus covers 20 languages, i.e. Arabic (Modern Standard Arabic), Bulgarian, Chinese (Mandarin and Shanghai), Croatian, Czech, French, German, Hausa, Japanese, Korean, Polish, Portuguese (Brazilian), Russian, Spanish (Latin American), Swedish, Tamil, Thai, Turkish, Ukrainian, and Vietnamese. This selection covers a broad variety of language peculiarities relevant for Speech and Language research and development. It comprises wide-spread languages (e.g. Arabic, Chinese, Spanish, Russian), contains economically and politically important languages, and spans wide geographical areas (Europe, Africa, America, Asia).

The spoken speech covers a broad selection of phonetic characteristics, e.g. tonal sounds (Mandarin, Shanghai, Thai, Vietnamese), pharyngeal sounds (Arabic), consonantal clusters (German), nasals (French, Portuguese), and palatized sounds (Russian). The written language contains all types of writing systems, i.e. logographic scripts (Chinese Hanzi and Japanese Kanji), phonographic segmental scripts (Roman, Cyrillic), phonographic consonantal scripts (Arabic), phonographic syllabic scripts (Japanese Kana, Thai), and phonographic featural scripts (Korean Hangul). The languages cover many morphological variations, e.g. agglutinative languages (Turkish, Korean), compounding languages (German), and also include scripts that completely lack word segmentation (Chinese, Thai).

2.2. Data Acquisition

The data acquisition was performed in countries where the language is officially spoken. In each language about 100 adult native speakers were asked to read about 100 sentences. The first batch of data collection was done from May 1996 to November 1997, and a second batch between 2003 and 2012. During the first batch we collected Arabic speech in Tunis, Sfax and Djerba, Tunisia; Mandarin in Beijing, Wuhan and Hekou, China; Shanghai in Shanghai, China; Croatian in Zagreb, Croatia, and parts of Bosnia; Czech in Prague, Czech Republic; French in Grenoble, France; German in Karlsruhe, Germany; Japanese in Tokyo, Japan; Korean in Seoul, Korea; Portuguese in Porto Velho and Sao Paulo, Brazil; Polish in Poland, Russian in Minsk, Belarus; Spanish in Heredia and San Jose, Costa Rica; Swedish in Stockholm and Vaernamo, Sweden; Tamil in India, and Turkish in Istanbul, Turkey. In the second batch we collected Bulgarian in Sofia, Hausa in Cameroon, Thai in Bangkok, Ukrainian in Donezk, and Vietnamese in Hanoi and Ho Chi Minh City.

The read texts were selected from national newspaper articles available from the web to cover a wide domain with large vocabulary. The articles report national and international political news, as well as economic news, which makes it possible to compare the usage of proper names (Politicians, companies, etc.) across languages. We used the following newspapers: Assabah for Arabic, Banker, Cash, and Sega for Bulgarian, Peoples Daily for Mandarin and Shanghai Chinese, HRT and Obzor Nacional for Croatian, Ceskomoravsky Profit Journal and Lidove Noviny newspaper for Czech, Le Monde for French, Frankfurter Allgemeine und Sueddeutsche Zeitung for German, CRI online and RFI for Hausa, Hankyoreh Daily News for Korean, Nikkei Shinbun for Japanese, Folha de Sao Paulo for Portuguese, Dziennik Polski for Polish, Ogonyok Gaseta and express-chronika for Russian, La Nacion for Spanish, Goeteborgs-Posten for Swedish, Thinaboomi Tamil Daily for Tamil, Bangkok Biz news and Daily News for Thai, Zaman for Turkish, Pravda among 9 other online newspapers for Ukrainian, and Tin Tuc among others for Vietnamese.

The speech data was recorded with a close-speaking microphone and is available in identical characteristics for all languages: PCM encoding, mono quality, 16bit quantization, and 16kHz sampling rate. Most recordings were done in ordinary rooms, in the majority without background noise, so that the speakers were not distracted. The quality of noise level and recording room setup is reported for each session. The speakers were given instructions about the equipment handling in advance. They were introduced to the projects goals and were allowed to read the texts before recording. The transcriptions are available in the original script of the corresponding language. In addition, all transcriptions have been romanized, i.e. transformed into Roman script applying reversible character mappings. The transcripts were internally validated and supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects such as breathing, laughing, and hesitations. Speaker information, such as age, gender, place of birth, dialect, occupation, etc. as well as information about the recording setup complement the database.

Table 1: GlobalPhone *Corpus Statistics*

Language	Training [hrs:min]	Development [hrs:min]	Evaluation [hrs:min]
Arabic	12:00	TBA	TBA
Bulgarian	16:47	2:16	1:56
Croatian	11:48	2:02	1:45
Czech	26:49	2:22	2:41
French	24:55	TBA	2:01
German	14:54	1:57	1:28
Hausa	6:36	1:02	1:06
Japanese	21:51	1:26	1:40
Korean	16:34	2:09	2:04
Mandarin	26:38	1:59	2:25
Portuguese	22:45	1:38	1:47
Polish	18:39	2:47	2:16
Russian	21:08	2:41	2:36
Shanghai	9:50	TBA	TBA
Spanish	17:35	1:40	2:03
Swedish	17:39	2:03	1:58
Tamil	15:50	1:04	1:00
Thai	19:05	2:03	1:58
Turkish	13:04	1:57	1:53
Ukrainian	11:32	1:13	1:07
Vietnamese	22:15	1:40	1:30
Total	368:14	33:59	35:14

2.3. Corpus Statistics

The entire GlobalPhone corpus contains over 400 hours of speech spoken by more than 2000 native adult speakers. The data are organized by languages and speakers and are divided into speaker disjoint sets for training (80%), development (10%), and evaluation (10%). Table 1 summarizes the amount of transcribed speech data per language.

3. Rapid Language Adaptation Toolkit (RLAT)

The project SPICE (NSF, 2004-2008) performed at the Language Technologies Institute at Carnegie Mellon and the Rapid Language Adaptation project at the Cognitive Systems Lab (CSL) aim at bridging the gap between the language and technology expertise. For this purpose RLAT [8] provides innovative methods and interactive web-based tools to enable users to develop speech processing models, to collect appropriate speech and text data to build these models, as well as to evaluate the results allowing for iterative improvements [9]. The toolkit significantly reduces the amount of time and effort involved in building speech processing systems for unsupported languages. In particular, the toolkit allows the user to (1) design databases for new languages at low cost by enabling users to record appropriate speech data along with transcriptions, (2) to continuously harvest, normalize, and process massive amounts of text data from the web, (3) to select appropriate phone sets for new languages efficiently, (4) to create vocabulary lists, (5) to automatically generate pronunciation dictionaries, (6) to apply these resources by developing acoustic and language models for speech recognition, (7) to develop models for text-to-speech synthesis, and (8) to finally integrate the built components into an application and evaluate the results using online speech recognition and synthesis in a talk-back function [9].

RLAT and SPICE leverage off the two projects GlobalPhone and FestVox [10] to implement bootstrapping techniques that are based on extensive knowledge and data sharing across languages, as well as sharing across system components [9]. Examples for data sharing techniques are the training of multilingual acoustic models across languages based on the definition of global phone sets. Sharing across components happens on all levels between recognition and synthesis, including phone sets, pronunciation dictionaries, acoustic models, and text resources.

RLAT [8] and SPICE are a freely available online service which provides an interface to the web-based tools and has been designed to accommodate all potential users, ranging from novices to experts. Novice users are able to read easy-to-follow, step-by-step guidelines as they build a speech processing system. Expert users can skip past these instructions. In addition, file-uploading routines allow for feeding the bootstrapping algorithms with available data and thus shortcut the process. As a result the tools collect information from the broadest array of people: a general audience of Internet users who may have little experience with speech tools, and a specific audience of speech and language experts, who can use data they already have. By keeping the users in the developmental loop, the RLAT and SPICE tools can learn from their expertise to constantly adapt and improve the resulting models and systems.

The tools are regularly used for training and teaching purposes at two universities (KIT and CMU). Students are asked to rely solely on the tools when building speech processing systems and report back on problems and limitations of the system. Results indicate that it is feasible to build end-to-end speech processing systems in various languages (more than 15) for small domains within the framework of a six-week hands-on lab course. Our tools will hopefully revolutionize the system development process in the future. Archiving the data gathered on-the-fly from many cooperative native users will significantly increase the repository of languages and resources. Data and components for under-supported languages will become available at large to let everyone participate in the information revolution, improve the mutual understanding, bridge language barriers, and thus foster educational and cultural exchange.

4. GlobalPhone Pronunciation Dictionaries

Phone-based pronunciation dictionaries are available for each GlobalPhone language. The dictionaries cover the words which appear in the transcriptions. The majority of the dictionaries were constructed in a rule-based manner using language specific phone sets. After this automatic creation process the dictionaries were manually post-processed word-by-word by native speakers, correcting errors in the automatic pronunciation generation and introducing pronunciation variants. To enable the development of multilingual speech processing, the phone names are consistent across languages, leveraging the International Phonetic Alphabet (IPA) [11]. Table 2 gives an overview of the size of the phone sets, amount of vocabulary words covered, and amount of pronunciation variants in the GlobalPhone pronunciation dictionaries.

Table 2: GlobalPhone *Pronunciation Dictionaries*

Languages	#Phones	#Words	#Dict entries
Bulgarian	44	275k	275k
Croatian	32	21k	23k
Czech	41	277k	277k
French	39	122k	195k
German	43	39k	41k
Hausa	33	43k	48k
Japanese	31	9k	13k
Korean	39	1.3k	3k
Mandarin	49	73k	73k
Portuguese	45	59k	59k
Polish	36	34k	34k
Russian	47	39k	40k
Spanish	42	31k	39k
Swedish	48	25k	25k
Tamil	41	288k	292k
Thai	44	23k	25k
Turkish	31	34k	34k
Ukrainian	51	40k	40k
Vietnamese	38	30k	39k

5. GlobalPhone Language Models

We applied RLAT to crawl a massive amount of text data and used the strategy presented in [12] to quickly and efficiently build the GlobalPhone language models for 19 languages. We crawled text data for several days, and each day one language model was built based on the daily crawled text data. The final language model was then created by a linear interpolation of all daily language models. The interpolation weights were computed using the SRI Language Model Toolkit [13], optimized on the GlobalPhone development sets. The experimental results in [12] indicated that the text data from the first few days are most helpful and therefore receive the highest interpolation weights in the final language model. Since the outcome of the crawling process depends on the input websites, the starting pages have to

be chosen carefully. In some cases (Croatian, Japanese, Korean, Thai) the crawling process finished prematurely. In those cases we selected additional websites to harvest more diverse text data. The final best language model were then built based on the interpolation of the language models from a variety of websites. Since some scripts lack a segmentation into words or do not provide a suitable definition of 'word units' (Chinese, Korean, Japanese, Tamil, Thai, and Vietnamese) we defined syllables or characters as token units for the purpose of speech recognition. Table 3 gives an overview of the amount of crawled text data, the trigram perplexities (PPL), out-of-vocabulary (OOV) rates, and the vocabulary sizes of the GlobalPhone language models, for both the full (LM) and the pruned benchmark language models (LM-BM), which are available for download from our website [7]. The symbols in parantheses after the language name indicate the token units used, i.e. (w) for word-based, (s) for syllable-based, and (c) for character-based token units.

Table 3: GlobalPhone Text Resources and Language Models

Language	3-gram PPL		OOV	#Vocab	#Tokens
	LM-BM	LM	[%]		
Bulgarian (w)	454	351	1.0	274k	405M
Croatian (w)	721	647	3.6	362k	331M
Czech (w)	1421	1361	4.0	267k	508M
French (w)	324	284	2.4	65k	-
German (w)	672	555	0.3	38k	20M
Hausa (w)	97	77	0.5	41k	15M
Japanese (s)	89	76	1.0	67k	1600M
Korean (c)	25	18	0	1.3k	500M
Mandarin (c)	262	163	0.8	13k	900M
Portuguese (w)	58	49	9.8	62k	11M
Polish (w)	951	904	0.8	243k	224M
Russian (w)	1310	1150	3.9	293k	334M
Spanish (w)	154	108	0.1	19k	12M
Swedish (w)	423	387	5.3	73k	211M
Tamil (s)	730	624	1.0	288k	91M
Thai (s)	70	65	0.1	22k	15M
Turkish (w)	-	45	13.2	29k	7M
Ukrainian (w)	594	373	0.5	40k	94M
Vietnamese (s)	218	176	0	30k	39M

6. Speech Recognition Systems

In this section, we present the large vocabulary speech recognition systems trained and evaluated on 19 GlobalPhone languages, i.e. all languages but Arabic and Shanghai, for the latter no transcripts are available at this point. For training, development, and evaluation we used the audio data as described in Table 1, the dictionaries shown in Table 2, and the language models listed in Table 3. All recognition systems were build in the same fashion. The systems use Bottle-Neck front-end features with a multilingual initialization scheme as proposed in [14]. In this approach a multilingual multilayer perceptron (ML-MLP) was trained using the training data from 12 languages (Bulgarian, Chinese, English, French, German, Croatian, Japanese, Korean, Polish, Russian, Spanish, and Thai). To initialize MLP training for a system, we select the output from the ML-MLP based on the IPA phone set and use it as a starting point for MLP training. All weights from the ML-MLP were taken and only the output biases from the selected targets were used.

To rapidly bootstrap the system, the phone models were seeded by the closest matches of the multilingual phone inventory MM7 [15] derived from an IPA-based phone mapping. The acoustic model uses a fully-continuous 3-state left-to-right Hidden-Markov-Model. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. For context-dependent acoustic models, we trained a quintphone system and stopped the decision tree splitting process at a specified language dependent threshold (varies between 500 and 3,000 leaves depending on the available amount of training data). After context clustering, a merge-and-split training was applied, which selects the number of Gaussians according to the amount of data. For all models, we used one global semi-tied covariance (STC) matrix after Linear Discriminant Analysis (LDA).

To model tonal languages such as Chinese, Hausa, Thai, and Vietnamese, we apply the "Data-driven tone modeling" approach, where all tonal variants of a phone share one base model. The information about the tone is added to the dictionary in form of a tone tag. These tags are used as questions to be asked in the context decision tree when building context-dependent acoustic models. This way, the data decide during model clustering whether different tonal variants of the same basic phone end up being represented by different models or share the basic phone model. Figure 1 illustrates the recognition performance for 19 GlobalPhone systems, tested on the evaluation set using both, the full language models (LM) and the pruned benchmark language models (LM-BM).

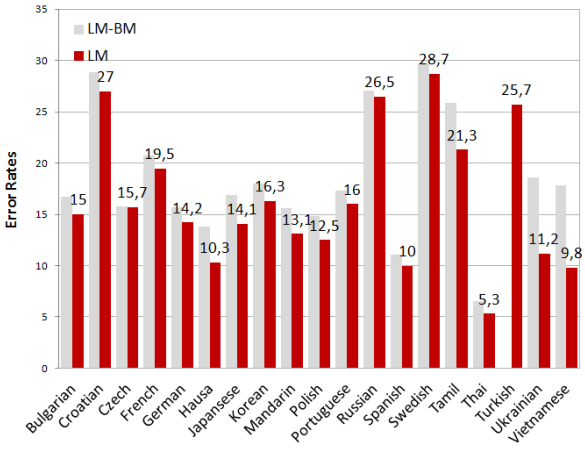

Figure 1: Word/Syllable/Character Error rates of the GlobalPhone speech recognition systems in 19 languages

7. Summary

In this paper we presented the latest status of the GlobalPhone speech and language resources in 20 different languages. We summarized the amount of speech data recordings, the number of entries covered in the pronunciation dictionaries and the amount of text data along with the characteristics of the language models. These resources are available to the community for research and development of multilingual speech processing systems. We also described the Rapid Language Adaptation Toolkit which was used to crawl additional text resources for language model building. Finally, we present the performance of our speech recognition systems based on the data described to provide a reference and benchmark numbers for researchers and developers who work with the GlobalPhone corpus.

8. References

[1] R. Gordon, Ed., *Ethnologue: Languages of the World*. Dallas: SIL International, 2005.

[2] T. Schultz, "Towards Rapid Language Portability of Speech Processing Systems," in *Conference on Speech and Language Systems for Human Communication (SPLASH)*, vol. 1, Delhi, India, November 2004.

[3] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*. Elsevier Academic Press, 2006.

[4] T. Schultz, "Globalphone: A Multilingual Speech and Text Database Developed at Karlsruhe University," in *Proceedings of the ICSLP*, 2002, pp. 345–348.

[5] ELRA, "European language resources association (ELRA)," ELRA catalogue. Retrieved November 30, 2012, from http://catalog.elra.info, 2012.

[6] Appen Butler Hill Pty Ltd, "Speech and Language Resources 2012," Appen Butler Hill Speech and Language Resources 2012 - Product Catalogue, 2012.

[7] LM-BM, "Benchmark GlobalPhone Language Models," Retrieved November 30, 2012, from http://csl.ira.uka.de/GlobalPhone, 2012.

[8] RLAT, "Rapid Language Adaptation Toolkit (RLAT)," Retrieved November 30 2012, from http://csl.ira.uka.de/rlat-dev, 2012.

[9] T. Schultz, A. W. Black, S. Badaskar, M. Hornyak, and J. Kominek, "Spice: Web-based tools for rapid language adaptation in speech processing systems," in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007.

[10] A. W. Black and K. Lenzo, "Building Voices in the Festival Speech Synthesis System," Festvox. Retrieved November 30 2012, from http://festvox.org/bsv/, 2000.

[11] IPA, *The principles of the International Phonetic Association*, 2nd ed. London, UK: University College of London, 1982.

[12] N. T. Vu, T. Schlippe, F. Kraus, and T. Schultz, "Rapid bootstrapping of five eastern european languages using the rapid language adaptation toolkit," in *INTERSPEECH*, 2010, pp. 865–868.

[13] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Intl. Conf. Spoken Language Processing (ICSLP)*, 2002.

[14] N. T. Vu, W. Breiter, F. Metze, and T. Schultz, "Initialization schemes for multilayer perceptron training and their impact on asr performance using multilingual data," in *INTERSPEECH*, 2012.

[15] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.