GLOBALPHONE: A MULTILINGUAL SPEECH AND TEXT DATABASE DEVELOPED AT KARLSRUHE UNIVERSITY

Tanja Schultz

Interactive Systems Laboratories
Karlsruhe University Germany, Carnegie Mellon University
E-mail: tanja@cs.cmu.edu

ABSTRACT

This paper describes the design, collection, and current status of the multilingual database *GlobalPhone,* an ongoing project since 1995 at Karlsruhe University. GlobalPhone is a high-quality read speech and text database in a large variety of languages which is suitable for the development of large vocabulary speech recognition systems in many languages. It has already been successfully applied to language independent and language adaptive speech recognition. GlobalPhone currently covers 15 languages Arabic, Chinese (Mandarin and Shanghai), Croatian, Czech, French, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. The corpus contains more than 300 hours of transcribed speech spoken by more than 1500 native, adult speakers and will soon be available from ELRA.

1 MOTIVATION

Nowadays, as the demand for rapid deployment of large vocabulary continuous speech recognition (LVCSR) in many languages grows, new approaches for cross-language transfer like the development of language independent global phone sets and language adaptive speech recognizers [1] are of increasing concern. These interests are accompanied by the need for a multilingual speech and text database that covers many languages and is uniform across languages. Uniformity here refers to the total amount of text and audio per language as well as to the quality of data, such as recording conditions (noise, channel, microphone etc.), collection scenario (task, setup, speaking style etc.), and transcription conventions. Only uniform data allow the development of global phone sets and enable the comparison of speech and/or text across languages. To train and evaluate LVCSR systems, dozens of hours of audio data from many speakers together with transcripts are required for acoustic modeling, and text data of millions of written words need to be available for language modeling. Furthermore, research on language independent and language adaptive speech recognition requires databases that cover the most relevant languages. Up to now no such uniform database which covers a large variety of languages was made available to the community.

2 CORPUS DESIGN

The aim of the GlobalPhone project was to provide a database to cope with the task of rapid deployment of LVCSR systems for native speakers in widespread languages. In order to achieve this goal, the objective was to collect at least 20 hours of transcribed speech (around 10,000 utterances or roughly 100,000 spoken words). To enable the development of reliable global phone sets, we targeted high-quality audio, such that the speech data mainly differ in the spoken language. It is our believe that global phone sets are crucial for rapid deployment but might also be useful for non-native speech recognition, however the latter was not the issue of this collection. A domain was chosen such that it is possible to additionally collect suitable large text corpora for language modeling by simply crawling the web.

2.1 Language Selection

About 4500 languages exist in the world, but the majority of languages are spoken by less than 100,000 speakers; only about 150 languages (3%) have more than 1 Million speakers.

Rank	*Language*	*Speakers official/native*		*Language Group*
1. *	Chinese	1000	907	Sino-Tibetan (Sinitic)
2.	English	1400	456	Indo-Euro (Germanic)
3.	Hindi	700	383	Indo-Euro (Indo-Iranian)
4. *	Spanish	280	362	Indo-Euro (Romance)
5. *	Russian	270	293	Indo-Euro (Slavic)
6. *	Arabic	170	208	Afro-Asiatic (Semitic)
7.	Bengalese	150	189	Indo-Euro (Indo-Iranian)
8. *	Portuguese	160	177	Indo-Euro (Romance)
9.	Malay	160	148	Austronesian (Polyn.)
10. *	Japanese	120	126	Isolate
11. *	French	220	123	Indo-Euro (Romance)
12. *	German	100	119	Indo-Euro (Germanic)
13.	Urdu	85	96	Indo-Euro (Indo-Iranian)
14.	Punjabi		89	Indo-Euro (Indo-Iranian)
15. *	Korean	60	73	Isolate

Table 1: Most widespread languages of the world and their pertaining speakers population (GlobalPhone languages are indicated by "*")

To select a representative subset of languages in a data collection for multilingual speech recognition, the following characteristics should be considered: (1) Size of speaker population, (2) Political and economic relevance, (3) Geographic coverage: European, Asian, Central Asian, Indian, African, American, as well as Minority languages, (4) Phonetic coverage, e.g. tones and pharyngeal sounds, (5) Orthographic script variety: phonologic scripts (e.g. alphabetic scripts like Latin, Cyrillic, Arabic), syllable-based scripts (e.g. Japanese Kana, Korean Hangul), and ideographic scripts (e.g. Chinese Hanzi and Japanese Kanji), (6) Morphologic variety: agglutinative languages (e.g. Turkish, Korean), compounding languages (e.g. German), and different word segmentation concepts (e.g. English, Chinese).

The GlobalPhone languages were selected following most but not all of these considerations; the size of speaker population and language relevance was favored above geographic coverage (no African languages so far). Some languages like Czech and Swedish were collected to study cross-language portability within language families. So far the following 15 languages have been collected: Arabic, Chinese (Mandarin and Shanghai), Croatian, Czech, French, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. Table 1 shows the ranking of the most widespread languages of the world derived from Webster's New Encyclopedia Dictionary 1992. Global-Phone languages are indicated by "*". Considering the fact that English is already available in a very similar framework (Wall Street Journal), the database covers 9 out of the 12 most frequent languages of the world.

2.2 Task and Text Selection

Since the most time and cost consuming process of a database collection is the transcription, we decided to collect speech data read from texts already electronically available. We selected widely read national newspapers available on the Internet as resources and chose texts from national and international political and economic topics to restrict the vocabulary. Most of the articles were collected between May 1996 and November 1997, which makes it possible to compare the usage of proper names, (e.g. Politician names, cities, companies, etc.) across languages. Furthermore, this approach allows collecting additional consistent text corpora for the training of statistical language models by crawling the web. Meanwhile huge newspaper text corpora are available from LDC and ELRA which enables the community to compare recognition results across sites.
For the GlobalPhone corpus, we used the following newspapers available on the Internet: Assabah for Arabic, Peoples Daily for Chinese, Hrvastka radiotelevizija (HRT) and Obzor Nacional for Croatian, Ceskomoravsky Profit Journal and Lidove Noviny newspaper for Czech, various sources from the ".fr" domain for French (see [2] for more details), Frankfurter Allgemeine and Süddeutsche Zeitung for German, Hankyoreh Daily News for Korean, Nikkei Shinbun for Japanese, Folha de São Paulo for Portuguese, Ogonyok Gaseta and express-chronika for Russian, La Nacion for Spanish, Göteborgs-Posten for Swedish, Thinaboomi Tamil Daily for Tamil, and Zaman for Turkish. Table 2 shows the according Internet links.

Language	*Internet link (status from April 2002)*
Arabic	http://www.assabah.press.ma
C-Mandarin	http://www.snweb.com.cn
C-Shanghai	http://www.snweb.com.cn
Croatian	http://www.hrt.hr/vijesti
	http://nacional.hr
	http://www.tel.hr.hrvatski-obzor
Czech	http://press.media.cz/press/pr/index.html
German	http://www.{faz,sueddeutsche}.de
Japanese	http://www.nikkeihome.co.jp
Korean	http://news.hani.co.kr
Portuguese	http://www.uol.com.br/fsp
Russian	http://www.ropnet.ru/ogonyok
Spanish	http://www.nacion.co.cr
Swedish	http://www.gp.se
Tamil	http://www.thinaboomi.com
Turkish	http://www.zaman.com.tr

Table 2: Newspapers used in GlobalPhone

3 DATA COLLECTION

Most of the GlobalPhone data was collected between May 1996 and November 1997. German, Swedish, and Tamil data was collected in 1998; Czech was collected at Charles University IFAL in 1999, and French at CLIPS-IMAG between 2000 and 2001 (see [2] for details).

3.1 Collection Sites

In order to avoid artifacts which might occur when collecting speech of native speakers living in a non-native environment, we collected all GlobalPhone data in the home countries of the native speakers: Modern Standard Arabic in Tunis and Sfax, 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; Brazilian Portuguese in Porto Velho and São Paulo, Brazil; Russian in Minsk, Belarus; Spanish in Heredia and San José, Costa Rica; Swedish in Stockholm and Vaernamo, Sweden; Tamil in South East India, and Turkish in Istanbul, Turkey.

3.2 Acquisition Session

For each language, about 100 native speakers were asked to read 20 minutes of text. The speech was recorded in one session per speaker. After the speakers were introduced to the project goals and equipment handling they were asked to read about 3-5 articles from the newspapers mentioned above. This corresponds to roughly 20 minutes of spoken speech per speaker. The speakers were allowed to read the text before recording in order to clarify exceptional words and their pronunciation and to minimize reading errors. Each session was recorded in one piece, but the speakers were instructed to pause at the end of every sentence during recording. This process made it easier to start over in case of reading errors or other problems, but still provided an easy solution to later segment the data into turns. In case of longer interruptions for questions or other issues, the recording device was halted.

All recordings were done in ordinary, but quiet rooms with only the subject and the coordinator present. The size of the rooms ranged from small to large and varied between office and private rooms; in very few cases, the recordings were done in public places. The surrounding noise level ranged from quiet to loud, however the majority of recordings were done in quiet environments, in order to not distract the speakers. The information about recording setup, environmental conditions and background noise level were documented for each session.

3.3 Recording Equipment

The recording tool consisted of a portable Sony DAT-recorder TDC-8 and a close-talking Sennheiser microphone HD-440-6. The data was digitally recorded at a 48kHz-sampling rate at 16bit linear quantization. For further processing, the data were downsampled to a 16kHz-sampling rate.

3.4 Data Validation and Completion

The recorded data was validated by a two-step approach. First, the downsampled DAT audio file of each speaker was split into turns by a silence detector. Since the speakers made a pause at the end of every sentence during recording, this was a relatively simple task. Second, the sentences of the text file were assigned to the turns. The same native experts who collected the data listened to the utterances and checked if the text corresponded to the speech. Clearly audible spontaneous effects like false starts, obvious hesitations and stuttering were marked, minor differences between text and speech were corrected, and incorrectly read utterances with major differences were deleted from the database.

4 CURRENT STATUS OF THE CORPUS

4.1 Subjects

The aim of the collection was to recruit equal numbers of subjects of both sexes, adult persons of various age categories and different education levels. In order to control the data proportion, we gathered demographic information from each speaker including (1) gender, (2) age, (3) place where the speaker was raised, (4) dominant dialect spoken by the speaker, (5) occupation (level of education), (6) health condition (cold or allergy), and (7) whether the speaker is a smoker or not.

The collected speakers' population contains almost equal numbers of subjects of both sexes, except for Japanese and German where the collection was done at computer science departments in which male students predominated. In all languages, students were the most dominant group among the subjects since they were easy to win for spending time and effort on giving speech. This fact is also reflected in figure 1 which shows the subjects' age distribution and demonstrates that the class of 20 to 29 year old subjects is the most frequent one.

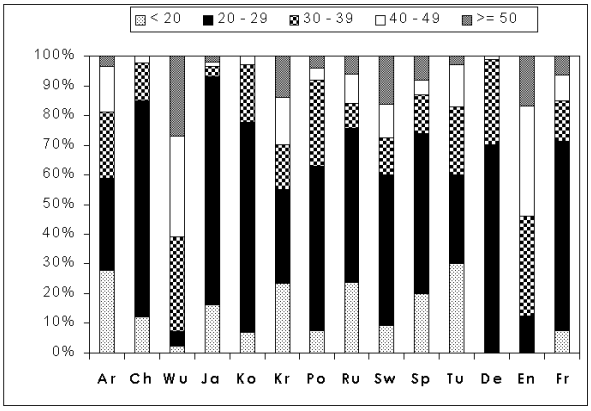

Figure 1: Speakers age distribution [Yrs]

4.2 Speech data

The format of all GlobalPhone speech data is PCM 16kHz, 16bit mono quality, high-byte first order without header. The files are losslessly compressed using the "Shorten", algorithm (version 1) by Tony Robinson, (http://www.softsound.com/Shorten.html). Table 3 shows the amount of audio data and number of speakers collected in each language. The overall amount of read speech data collected in 15 languages is about 328 hours spoken by 1506 native speakers. The average length per turn is about 9sec. The average number of words spoken in a turn is about 19 units, but varies across languages with the length of the word unit (segmentation).

Language	*Number Speakers*	*Audio [hours]*	*Spoken Words*
Arabic	170	35	i.p.
Ch-Mandarin	132	31	263k
Ch-Shanhai	41	10	95k
Croatian	92	16	120k
Czech	102	29	220k
French	94	25	250k
German	77	18	151k
Japanese	144	34	268k
Korean	100	21	117k
Portuguese	101	26	208k
Russian	106	22	170k
Spanish	100	22	172k
Swedish	98	22	184k
Tamil	49	i.p.	i.p.
Turkish	100	17	113k
Total	1506	328	2331k

Table3: Current status of the GlobalPhone corpus (i.p. = in progress)

4.3 Text data

The transcripts are available in the original orthographic script, but were additionally mapped into a romanized form. For Arabic only the romanized version is available; Tamil is not processed yet. For romanization, several tools were developed which vary from simple context-free mapping tools to more elaborated algorithms, like for the segmentation and pinyinzation of Chinese Hanzi. The romanized version of all transcripts is coded in ASCII-7. The original orthographic scripts are coded in various different formats since we followed the encodings that were provided by the Internet sources. The 1-byte coded orthographic scripts use ISO8859-1 for French, German, Portuguese, Spanish, and Swedish, ISO8859-2 for Croatian and Czech, ISO8859-9 for Turkish, and KOI8 for Russian. The 2-byte coded scripts use JIS coding for Japanese Kanji, Johabsh coding for Korean Hangul, and Guobiao coding for Chinese Hanzi.

4.4 Data Partition

For each language, the data was divided into 3 sets, one set for training, one set for cross-validation, and one evaluation set for reporting final numbers. The sets were split up at an 80:10:10 ratio in such a way that no speaker appears in more than one group and no article is read by two speakers from different groups.

CONCLUSIONS

The GlobalPhone corpus provides read speech data in the most widespread languages of the world. As a result of the uniform collection procedure, it is suitable for language independent and language adaptive large vocabulary continuous speech recognition as well as for language identification tasks. The entire GlobalPhone corpus enables the acquisition of acoustic-phonetic knowledge of the following 15 spoken languages Arabic, Chinese-Mandarin, Chinese-Shanghai, Croatian, Czech, French, German, Japanese, Korean, Portuguese (Brazilian), Russian, Spanish, Swedish, Tamil, and Turkish. In each language 100 speakers read each about 100 sentences. The read texts were selected from national newspapers available via Internet to provide a large vocabulary (up to 65,000 words). The read articles cover national and international political news as well as economic news from 1995-2000. The speech is available in 16bit, 16kHz mono quality, recorded with a close-speaking microphone. The transcriptions are internally validated and supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects like laughing and hesitations. Speaker information such as age, gender, occupation, etc. as well as information about the recording setup complement the database. The entire GlobalPhone corpus contains over 300 hours of speech spoken by more than 1500 native, adult speakers. The database has already been successfully applied to language independent and language adaptive speech recognition as well as to speaker, accent, and language identification.

ACKNOWLEDGEMENTS

The author gratefully acknowledges all members of the GlobalPhone team for their great enthusiasm during the data collection and evaluation: Omar Abdallah, Jamal Abu-Alwan, Hiroko Akatsu, Giovanni Najero Barquero, Kenan Çarkı, Keal-Chun Cho, Caleb Everett, Raul Ivo Faller, Renato Ferreira, Sanela Habibija, Wajdi Halabi, Evelyn Kimmich, Kyung-Kyu Lee, Natalia and Orest Mikhailiouk, Jae-Ho Park, Martin Sjögren, Sang-Hun Shin, Maho Takeda, Sayoko Takeda, Jing Wang, Tianshi Wei, Jiaxing Weng, Mutlu Yalçın , Nadia Zouabi, Olfa Karboul-Zouari, and Mohammed Zouari. Special thanks to my former colleagues Thomas Kemp for collecting the German part of GlobalPhone and Laura Mayfield-Tomokiyo for the Japanese data collection.

REFERENCES

[1] T. Schultz and A. Waibel, *Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition*, Speech Communication, Volume 35, Issue 1-2, pp 31-51, August 2001.

[2] D. Vaufreydaz, C. Bergamini, J.F. Serignat, L. Besacier, and M. Akbar, *A New Methodology for Speech Corpora Definition from Internet Documents,* Proceedings of the LREC 2000, Athens, Greece 2000.

[3] European Language Resources Association (ELRA): http://www.icp.grenet.fr/ELRA/home.html