The GlobalPhone Project: Multilingual LVCSR with JANUS-3

Tanja Schultz, Martin Westphal, and Alex Waibel
{tanja,westphal,ahw}@ira.uka.de

Interactive Systems Laboratories
University of Karlsruhe and Carnegie Mellon University
Am Fasanengarten 5, D-76131 Karlsruhe, GERMANY
Tel: ++49 721 608 4735 Fax: ++49 721 607721

Abstract. This paper describes our recent effort in developing the GlobalPhone database for multilingual large vocabulary continuous speech recognition. In particular we present the current status of the GlobalPhone corpus containing high quality speech data for the 9 languages Arabic, Chinese, Croatic, Japanese, Korean, Portuguese, Russian, Spanish, and Turkish. We also discuss the JANUS-3 toolkit and how it can be applied on our way towards multilinguality using the GlobalPhone database.

1 Introduction

The increasing interest in large vocabulary multilingual speech recognition and understanding systems over the last years is accompanied by the development of multilingual databases like the OGI corpus and the CALL HOME task. These databases are collected over telephone lines, which makes the speaker acquisition simple and fast.

For training and testing large vocabulary speaker-independent speech recognition systems a large amount of transcribed speech data is necessary. To achieve robust acoustic models about 3000 utterances (around 80.000 spoken words) are sufficient, but for training n-gram language models more than 7000 utterances should be available. Transcribing of conversational speech is one of the most expensive and time consuming step of a database collection.

Now, as the demand for portability and fast development of recognition systems in several languages grows, techniques for rapid cross-language transfer like bootstrapping from multilingual phoneme sets is of increasing concern. The development of reliable multilingual phoneme sets and the evaluation of these techniques requires high quality speech data, which guarantees that the only difference of the acoustic data is the spoken language. As could be seen in [1] different quality conditions can influence the language identification significantly.

The paper is organized as follows: In the first section we describe the corpus design and the collection scenario of the GlobalPhone database. In the second section the current status of the corpus is reported. The third section discusses the design principles of JANUS-3 and presents the recognition performance on various tasks in several languages.

2 The GlobalPhone Data Collection

We have collected a multilingual high quality speech corpus called GlobalPhone, which is suitable for the development of multilingual large vocabulary continuous speech recognition systems.

2.1 Corpus Design

Language Selection More than 6000 different languages exist in the world, about 195 are spoken by more than 1 million people. The languages to start the GlobalPhone project with are selected based on their variability and distribution, the importance in the world, and on linguistic considerations. The table below shows the ranking of the most widespread languages of the world [2]. For the present the GlobalPhone database consists of 9 different languages namely Arabic (Tunisia), Chinese (Mandarin), Croatic (Croatia and Bosnia), Japanese, Korean, Portuguese (Brazil), Russian (Belorussia), Spanish (Costa Rica), and Turkish. Considering the fact that English, French, and German are already available in similar frameworks, we are now able to cover 9 out of the 12 most frequent languages of the world.

Language	Primary Locales	Speaker Language Group
1. Mandarin	China	907 Mio Sino-Tibetan (Sinitic)
2. English	USA, UK, Can, Australia	456 Mio Indo-European (Germanic)
3. Hindi	India	383 Mio Indo-European (Indo-Irania)
4. Spanish	Latin-America, Spain	362 Mio Indo-European (Romance)
5. Russian	Russia, Indep. States	293 Mio Indo-European (Slavic)
6. Arabic	N. Africa, Mid East	208 Mio Afro-Asiatic (Semitic)
7. Bengali	Bangladesh, India	189 Mio Indo-European (Indo-Irania)
8. Portuguese	Brazil, Portugal, Angola	177 Mio Indo-European (Romance)
9. Malay-Indo.	Indonesia, Malay, Brunei	148 Mio Austronesian (Polynesien)
10. Japanese	Japan	126 Mio Isolate
11. French	F, Can, Africa, Schwitzerl.	123 Mio Indo-European (Romance)
12. German	G, Austria, Schwitzerl.	119 Mio Indo-European (Germanic)
15. Korean	Korea, China	73 Mio Isolate
21. Italian	Italy, Schwitzerland	63 Mio Indo-European (Romanic)
25. Turkish	Turkey	57 Mio Altaic (Turkik)
44. Serbo-Croatian	Balkan Europe	20 Mio Indo-European (Slavic)

Table 1. Languages of the world

Task Selection The most time and cost consuming process of a database collection is the transcription of speech, therefore we decided to collect speech data read from text which is already electronic available. The collection scenario ensures, that vocabulary and task are similar across all languages. To achieve reasonable vocabulary coverage with acceptable OOV-rates we prefer a domain limited task. Additionally our intention was to include cross lingual words like proper names and products. These constraints were fulfilled by collecting texts of national newspapers available via Internet with national and international political and economic topics (similar to Wall Street Journal). Furthermore, this approach allows us to collect additional consistent text data for the training of n-gram language models.

For the GlobalPhone corpus we use *Assabah* for Arabic, *Peoples Daily* for Chinese, *HRT* and *Obzor Nacional* for Croatian, *HanaBBS* for Korean, *Nikkei Shinbun* for Japanese, *Folha de Sao Paulo* for Portuguese, *Ogonyok Gaseta* and *express-chronika* for Russian, *La Nacion* for Spanish, and *Zaman* for Turkish.

2.2 Data Acquisition

How to get non telephone speech data of native speaker? The most cost-saving way is to find foreign speakers living in the area, but we intended to avoid artefacts which might occur when collecting speech of native speaker living in a non native environment. The only alternative is to collect the database by native speakers in their home country, where the language is spoken. We employed foreign students at our university and taught them about the goals of our project and the details about the recording session. They learned how to handle the recording equipment and how to give useful instructions to the donators. The students prepared the text to be read and organized the subject recruitment before going to their home country.

Subject Recruitment We intended to collect equal numbers of subjects of both sexes, in all age categories and education levels. The subject population consists of 365 male and 409 female native speakers between the ages of 18 and 81. In each language about 100 native speakers were asked to read 20 minutes of the prepared newspaper texts. They were allowed to read the text before recording.

Data sheet Design A data sheet was designed to describe the recording session for each speaker. It contains information about the recording setup, the speaker characteristics, and the environmental conditions. In particular, we asked for recording duration, date and time as well as the identification and order of the articles. Details about the speaker consist of sex, age, occupation, dialect and health conditions. Questions of environmental setup include room characteristics and classification of the background noise level.

Recording Equipment The speech was recorded in office quality with a portable Sony DAT-recorder TDC-8 and a close-speaking Sennheiser microphone HD-440-6. The data was digitally recorded at 48 kHz sampling rate in stereo quality. For further processing the data files were transfered to a UNIX workstation, and sampled down to 16kHz with 16bit resolution in mono quality.

2.3 Data Validation

For the still ongoing validation process a special tool was created based on the JANUS-3 toolkit. The following three-pass approach was applied to process the data. First the data sheets were fed into the workstation. Based on this information a file for each speaker was created containing the relevant information. Second the transfered speech audio file of each speaker was splitted into chunks by a silence detector. The speakers were instructed to pause at the end of every sentence during recording. So the silence energy threshold and minimum duration of silence could be adjusted that each chunk roughly corresponds to one sentence. Third the sentences of the text file were assigned to the chunks. The same native experts who collected the data listened to the utterances and checked if the text corresponds to the speech. Clearly audible spontaneous effects like false starts, obvious hesitations and stuttering were marked, minor differences between text and speech were corrected, incorrect read utterances with major differences deleted.

3 Current Status of the Corpus

The GlobalPhone database currently consists of 9 different languages. Table 2 shows the status of the GlobalPhone database. The corpus size for each language is in the order of the English Wall Street Journal task (WSJ0) and the French BREF task. We assume that the collected speech data is sufficient for the training of robust acoustic models and language models. To calculate better n-gram models we have collected text data of the described newspapers via Internet.

Language	Speaker	Distinct Articles	Spoken Words	Rec. Hours	Vocab Size	Valid. Utts	Average Length	Words per Utt
Arabic	93	400	-	28	-	-	-	-
Chinese[1]	132	430	125K	40	4K	1008	10.5 sec	29.5
Croatic	85	131	89K	18	17K	2638	13.0 sec	27.5
Japanese[1]	121	534	182K	41	21K	2043	9.0 sec	21.4
Korean	70	210	-	32	-	-	-	-
Portuguese	75	97	182K	33	6K	182	9.3 sec	17.6
Russian	99	104	186K	26	20K	1956	6.8 sec	15.4
Spanish	89	277	164K	20	21K	1023	10.8 sec	27.3
Turkish	100	223	110K	18	15K	4120	8.9 sec	16.5

Table 2. The GlobalPhone Database

[1] For the Chinese hanse segmentation we implemented a new tool, for the Japanese kanji and kana we used CHASEN

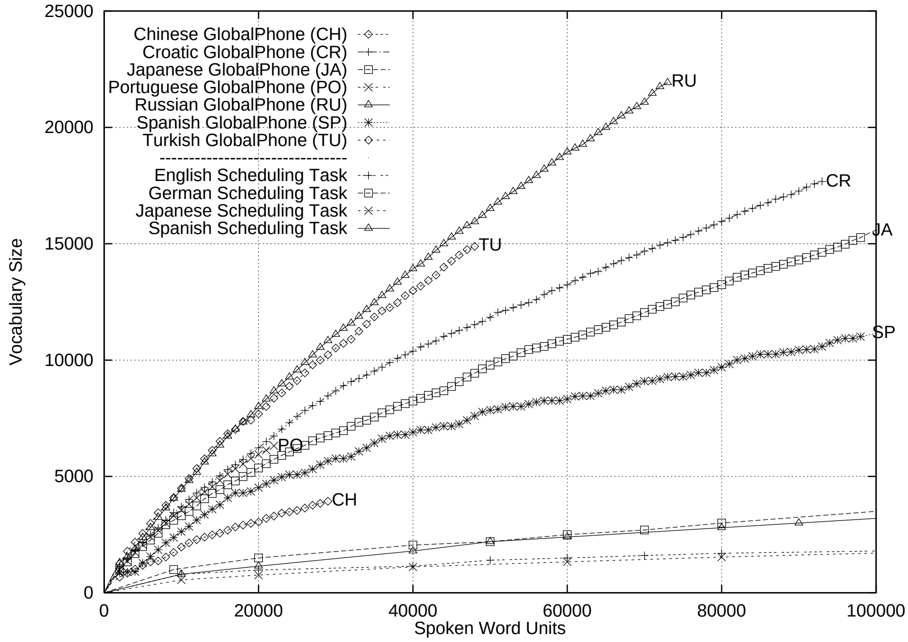

Fig. 1. Growth of Vocabulary Size

On average we have 97 speakers per language, however the number of speakers varies between 70 and 132. The definition of the unit "word" is not comparable for all languages. The Korean, Chinese and Japanese language does not have a word concept in the general meaning. Since there is no delimitation between written words the phrases have to be segmented into word units, requiring morphological analysis. The segmentation approach we applied results in relatively small vocabulary growth rates as can be seen from figure 1. In opposite Croatic, Russian, and Turkish have definite word boundaries, but these languages are highly inflecting. Since we count each flexion as one word we obtain very high growth rates. Figure 1 compares the growth rates of the read speech corpus GlobalPhone with the conversational speech data of the Spontaneous appointment Scheduling Task (SST). The latter is a very limited domain task which leads to very good coverage and small vocabulary size with low growth rates.

4 Multilingual speech recognition

JANUS is a speech-to-speech translation research project [3]. Speech translation consists of multiple parts (recognition, analysis, generation, and synthesis). Since we do not address anything beyond recognition in this paper, we will use the term "JANUS-3" meaning the latest version of the project's speech recognizer.

4.1 The JANUS-3 Speech Recognition Toolkit

As can be seen in table 3 JANUS-3 was recently used on various speech tasks that cover spontaneous and read speech as well as several languages [4]. Since we used JANUS-3 on the Japanese language we gained experience that enabled us to develop a well performing Japanese speech recognition system for GlobalPhone within an acceptable amount of time. One focus of the GlobalPhone project is to achieve rapid prototyping on new languages and find techniques to bootstrap new systems in an effective manner [5].

language	Task	Word error
English	Wall Street Journal	7.8%
English	Switchboard	36%
English	Scheduling Task	23%
German	Scheduling Task	14%
Spanish	Scheduling Task	17%
Japanese [2]	Scheduling Task	10%
Japanese [3]	GlobalPhone	16%

Table 3. Performance of the JANUS-3 system

4.2 The Design of JANUS-3

The original purpose of JANUS-3 was to create a powerful tool for speech recognition research. This means, that in many cases flexibility, portability, and expandability was considered to be more important than speed or memory consumption. Although JANUS-3's primary design goal was not efficiency, there are many ways to speed up and actually build a near realtime demo system with only small losses in recognition accuracy. Before JANUS-3 was written, the JANUS team made a lot of experiences with its predecessor JANUS-2. The ideas that were collected during this time were incorporated into the JANUS-3 design goals:

- **Flexibility**: JANUS-3 should be able to run many different speech recognizer architectures, different kinds of HMMs (discrete, semi-continuous, continuous), different kinds of emission probability computation (Gaussian mixtures, neural nets), and arbitrary sets of acoustic models (polyphones, clustering, subphones, phone-tags).
- **Factory-Style**: It should be possible to build a simple recognizer using a predefined development scheme by declaring the architecture and pushing a single button. This means all the development steps should fit into a single

[2] calculated based on hiragana word level
[3] calculated based on kanji single word

script that can run from the very beginning to the very end without any user interaction.

– **Foolproofness**: When doing research, people often make mistakes that can cost them a lot of time. JANUS-3 should offer many ways to make sure that everything is doing fine, and it should allow the user to control many of its functions.
– **Easy Usage**: It should be possible for the user to train a recognizer from different degrees of insight. The simplest way would be the one-button factory-style recognizer. If a user would like to do some non-standard things, or perform some experiments, then this should be easily possible, as well as accessing and manipulating the deepest and most nested data structures. To do all this, a script language is necessary, which allows close interaction with C and enables high level constructs. We chose Tcl[1] for that purpose.
– **Appearance**: JANUS-3 should "look nice". The Tcl/Tk language[2] makes programming of graphical user interfaces easy. Thus looking at datastructures and controlling the recognizer becomes simple.
– **Expandability**: JANUS-3 should be built very modular, such that it is easy to add new modules, and to use existing modules to build new ones upon them. Also, the object-oriented JANUS-3-programming-language (Tcl augmented by the JANUS-3-defined features) allows to design arbitrary recognizer architectures easily.

4.3 Towards Multilinguality using the GlobalPhone Corpus

Due to its design principles JANUS-3 is predestined to be a teaching tool. In a practical class we used it to explain students how state-of-the-art speech recognition systems are trained and tested. In several sessions they learned Tcl and the basic handling of JANUS-3 objects. They did experiments using the preprocessing, training and test modules. We demonstrate the design principle 'Appearance' of JANUS-3 by illustrating the results through graphical interfaces. Audio and preprocessed signals for example can be visualized with 'featview' and viterbi paths can be analysed with 'pathview' as shown in Figure 2. An introduction about language model tools completed the general part of their practical training.

In the language specific part the students were now able to train the speech recognizer in several languages using the validated parts of the GlobalPhone corpus as training and test database. First results are very promising although they are based on a very simple approach as well as insufficient data and thus can not be compared with our state-of-the-art systems. The experiments confirm the portability of JANUS-3 to other languages.

After finishing the validation of all the collected speech we will have sufficient data for the training of acoustic and language models. This will allow us to develop reasonable language dependent recognition systems, which we intend to do in a general language independent JANUS-3 approach.

---

[1] Tcl means Tool command language

[2] Tk is a toolkit extension for Tcl

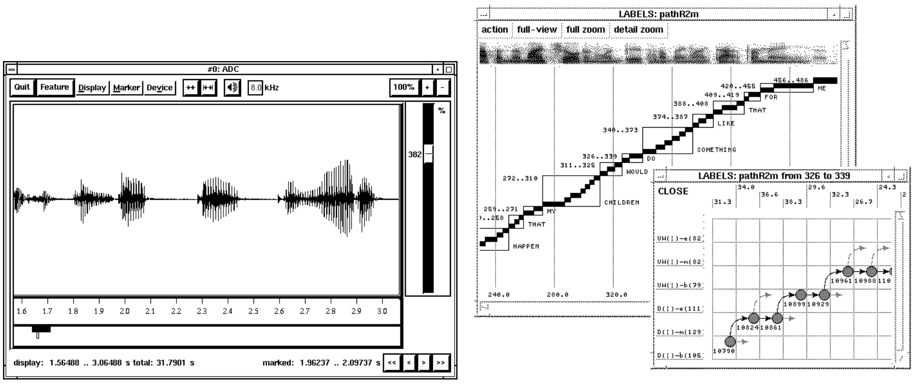

Fig. 2. left: 'featview' shows audio signals and preprocessed features;
right: 'pathview' shows viterbi alignments

5 Acknowledgement

Thanks to all members of the GlobalPhone team: Olfa Karboul Zouari and Mohamed Zouari (Arabic), Tianshi Wei, Jing Wang, and Jiaxing Weng (Chinese), Sanela Habibija (Croatic), Laura J. Tomokyo, Hiroko Akatsu, and Sayoko Takeda (Japanese), Keal-Chun Cho (Korean), Orest and Natalia Mikhailiouk (Russian), Raul Ivo Faller and Caleb Everett (Portuguese), Giovanni Najera Barquero (Spanish), and Mutlu Yalcin (Turkish). This research would not have been possible without their great enthusiasm during collection and validation of the database.

References

1. T. Schultz, I. Rogina, and A. Waibel: *LVCSR-based Language Identification.* Proceedings of the ICASSP 96, Atlanta, USA, May 1996.
2. *Webster's New Encyclopedic Dictionary.* Black Dog & Leventhal 1992.
3. A. Lavie et al.:*JANUS-III: Speech-to-Speech Translation in Multiple Languages.* To appear in ICASSP 1997.
4. B. Suhm et al.: *JANUS: Towards Multilingual Spoken Language Translation.* DARPA Speech and Natural Language Workshop, 1995.
5. T. Schultz and A. Waibel: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme sets.* Submitted to Eurospeech 1997.

This article was processed using the LaTeX macro package with LLNCS style