NON-NATIVE SPEECH DATABASES

Martin Raab[1,2]*, Rainer Gruhn*[1,3]*, Elmar Noeth*[2]

[1]Harman Becker Automotive Systems, Speech Dialog Systems, Ulm, Germany
[2]University of Erlangen-Nuremberg, Dept. of Pattern Recognition, Erlangen, Germany
[3]University of Ulm, Dept. of Information Technology, Ulm, Germany
mraab@harmanbecker.com

ABSTRACT

This paper presents a review of already collected non-native speech databases. Although the number of non-native speech databases is significantly less than the one of common speech databases, there were already a lot of data collection efforts taken at different institutes and companies. Because of the comparably small size of the databases, many of them are not available through the common distributors of speech corpora like ELDA or LDC. This leads to the fact that it is hard to keep an overview of what kind of databases have already been collected, and for what purposes there are still no collections. With this paper we hope to provide a useful resource regarding this issue.

Index Terms— non-native, multilingual, speech recognition, corpora, databases

1. INTRODUCTION

Recognition of non-native speech is getting more and more important for speech recognition systems. Current speech recognition system still show severe performance losses when confronted with it. Due to the intrinsic attributes of non-native speech, it is rather unlikely to solve this problem by a common statistical approach. The deviations of non-native speech from native speech are too severe, too unpredictable, and they depend on a couple of influences like native language and proficiency of the speaker in the foreign language.

These attributes make it impossible to collect enough training data for all possible combinations of speakers of one native language speaking a foreign language. Yet, previous work has shown that large improvements can be gained even with little non-native adaptation data [1, 2]. Hence it is valuable for developers of commercial speech recognition systems to be aware what non-native speech databases are already collected.

Furthermore, for all techniques that try to improve non-native speech recognition, appropriate non-native testing data is needed. The table of databases presented in the next sections should be a helpful tool to find suitable test sets.

This paper is organized as follows. Section 2 presents the table of non-native databases the authors of this paper have been able to identify. Section 3 briefly introduces some of the larger and more promising databases that have been found. Finally a conclusion is drawn about the current status of corpora for non-native speech recognition.

2. NON-NATIVE DATABASES

2.1. General Information

There is no claim of the authors for the following table to contain all existing databases. For example, there are some meeting/presentation databases that are not contained. These special databases are regarded as being less relevant, as the speech they contain is difficult due to a variety of different influences what makes it hard to concentrate solely on non-native speech recognition. Apart from these databases, the table provides details about all databases the authors were able to identify.

Table 2 gives the following information about each corpus: The name of the corpus, the institution where the corpus can be obtained, or at least further information should be available, the language which was actually spoken by the speakers, the number of speakers, the native language of the speakers, the total amount of non-native utterances the corpus contains, the duration in hours of the non-native part, the date of the first public reference to this corpus, some free text highlighting special aspects of this database and a reference to another publication. The reference in the last field is in most cases to the paper which is especially devoted to describe this corpus by the original collectors. In some cases it was not possible to identify such a paper. In these cases the paper where the authors first found information about this corpus is referenced.

Some entries are left blank and others are marked with unknown. The difference here is that blank entries refer to attributes the authors of this paper were not able to find information about. Unknown entries, however, indicate that no information about this attribute is available in the database itself. As an example, in the Jupiter weather database [3] no in-

Arabic	A	Italian	I
Chinese	C	Japanese	J
Czech	Cze	Korean	K
Danish	D	Malaysian	M
Dutch	Dut	Norway	N
English	E	Portugese	P
French	F	Russian	R
German	G	Spanish	S
Greek	Gre	Swedish	Swe
Hebrew	H	Thai	T
Indonesian	Ind	Vietnamese	V

Table 1. Abbreviations for languages used in Table 2

formation about the origin of the speakers is given. Therefore this data would be less useful for verifying accent detection or similar issues.

Where possible, the name is a standard name of the corpus, for some of the smaller corpora, however, there was no established name and hence an identifier had to be created. In such cases, we chose a combination of the institution and the collector of the database.

In the case where the databases contain native and non-native speech, the authors tried to only list attributes of the non-native part of the corpus. Most of the corpora are collections of read speech. If the corpus instead consists either partly or completely of spontaneous utterances, this is mentioned in the *Specials* column.

2.2. Glossary

In the table of non-native databases some abbreviations for language names are used. They are listed in Table 1. The actual table with information about the different databases is shown in Table 2.

2.3. Updates

We also created a companion website for this paper at wikipedia [4]. We invite everybody to add missing information for the benefit of all interested researchers.

3. MORE DETAILS FOR SOME DATABASES

3.1. CSLU

This database is special because of its wide variety of non-native accents (22 non-native accents). The corpus contains utterances of foreign speakers that were asked to speak about themselves in English for 20 seconds. The recordings were conducted over telephone. A strength of this corpus is that it contains human judgments of the level of foreign accent on a level from one to four. 5000 utterances with each 20 seconds result in a estimated corpus size of almost 30 hours.

This corpus is available through LDC.

3.2. Cross Towns

The strength of this corpus is that it covers many language directions (speakers of one native language speaking another language). Altogether the corpus cover 24 different language directions. Each recording of a language direction contains two times 45 city names per speaker. First the 45 city names are read from a prompt, and then they are repeated after listening to the name via headphone. 13000 of the utterances are manually transcribed at the phonetic level, and there is information about the language proficiency of the speakers.

A planned release at ELRA in 2006 did not succeed. According to the author of the corpus a future release of this corpus is undetermined.

3.3. NATO HIWIRE

This corpus contains speech from 81 different speakers. The utterances were collected in a military pilot-ground control conversation task. An advantage for general speech recognition of this corpus is, that is was originally collected in a studio and only later convolved with typical cockpit noise. As a consequence, the corpus now contains two signal levels, one clean speech and and one with added noise.

For information about obtaining this corpus contact one of the authors of [5].

3.4. CLIPS-IMAG

This corpus differs from other corpora because of its untypical combination of languages. Whereas most databases focus on English as foreign language, this corpus contains French non-native speech by Vietnamese and Chinese speakers. With a total amount of 6h of non-native speech this corpus is also relatively large. The speech covers dialogues and articles from the tourist domain. Although this makes the nature of the sentences spontaneous, the data is read speech.

For information if and how this corpus can be obtained we recommend contacting the authors of [6].

3.5. ATR-Gruhn

Another corpus which has the advantage of proficiency ratings of the speakers. The corpus contains 89 non-native speakers of English, their origin is evenly distributed between speakers from China, France, Germany, Indonesia and Japan. Additionally, there are seven more speakers with other native languages.

Each speaker read 25 credit card number sequences, 48 phonetically rich sentences and six hotel reservation dialogs. The proficiency rating was performed by native English speakers from the US and Canada with teaching experience.

This database is available at ATR (or its successor NICT).

Table of Databases

Corpus	Author	Available at	Language(s)	#Speakers	native Language	#Utt.	Duration	Date	Specials	Reference
ATR-Gruhn	Gruhn	ATR	E	96	C G F J Ind	15000		2004	proficiency rating	[7]
BAS Strange Corpus I+II		ELRA	G	139	50 countries	7500		1998		[8]
Broadcast News		LDC	E					1997		[2]
Berkeley Restaurant		ICSI	E	55	G I H C F S J	2500		1994		[9]
Cambridge-Witt	Witt	U. Cambridge	E	10	J I K S	1200		1999		[1]
Cambridge-Ye	Ye	U. Cambridge	E	20	C	1600		2005		[10]
Children News	Tomokiyo	CMU	E	62	J C	7500		2000	partly spontaneous	[2]
CLIPS-IMAG	Tan	CLIPS-IMAG	F	15	C V		6h	2006		[6]
CLSU		LDC	E		22 countries	5000		2007	telephone, spontaneous	[11]
CMU		CMU	E	64	G	452	0.9h		not available	[12]
Cross Towns	Schaden	U. Bochum	E F G I Cze Dut	161	E F G I S	72000	133h	2006	city names	[13]
Duke-Arslan	Arslan	Duke University	E	93	15 countries	2200		1995	partly telephone speech	[14]
ERJ	Minematsu	U. Tokyo	E	200	J	68000		2002	proficiency rating	[15]
Fitt	Fitt	U. Edinburgh	F I N Gre	10	E	700		1995	city names	[16]
Fraenki		U. Erlangen	E	19	G	2148				[17]
Hispanic	Byrne		E	22	S		20h	1998	partly spontaneous	[18]
IBM-Fischer		IBM	E	40	S F G I	2000		2002	digits	[19]
ISLE	Atwell	EU/ELDA	E	46	G I	4000	18h	2000		[20]
Jupiter	Zue	MIT	E	unknown	unknown	5146		1999	telephone speech	[3]
LDC WSJ1		LDC		10		800	1h	1994		[2]
MIST		ELRA	E F G	75	Dut	2200		1996		[21]
NATO HIWIRE		NATO	E	81	F Gre I S	8100		2007	clean speech	[5]
NATO M-ATC	Pigeon	NATO	E	622	F G I S	9833	17h	2007	heavy background noise	[22]
NATO N4		NATO	E	115	unknown		7.5h	2006	heavy background noise	[23]
Onomastica			D Dut E F G Gre I N P S Swe			(121000)		1995	**only lexicon**	[24]
PF-STAR		U. Erlangen	E	57	G	4627	3.4h	2005	children speech	[25]
Sunstar		EU	E	100	G S I P D	40000		1992	parliament speech	[26]
TC-STAR	Heuvel	ELDA	E S	unknown	EU countries		13h	2006	multiple data sets	[27]
TED	Lamel	ELDA	E	40(188)	many		10h(47h)	1994	eurospeech 93	[28]
TLTS		DARPA	A		E		1h	2004		[29]
Tokyo-Kikuko		U. Tokyo	J	140	10 countries	35000		2004	proficiency rating	[30]
Verbmobil		U. Munich	E	44	G		1.5h	1994	very spontaneous	[31]
VODIS		EU	F G	178	F G	2500		1998	about car navigation	[32]
WP Arabic	Rocca	LDC	A	35	E	800	1h	2002		[33]
WP Russian	Rocca	LDC	R	26	E	2500	2h	2003		[34]
WP Spanish	Morgan	LDC	S		E			2006		[35]
WSJ Spoke			E	10	unknown	800		1993		[36]

Table 2. Overview of non-native Databases

3.6. ISLE

ISLE is one of the largest corpora (measured in hours) and has the advantage to be distributed by ELDA for a moderate price. There are only two accents, German and Italian accented English in this corpus. The speakers read 1300 words of a non-fictional, autobiographic text and 1100 words in short utterances which were designed to cover typical pronunciation errors of language learners. The corpus is annotated at the word and at the phone level, which makes it especially interesting for the development of Computer Assisted Language Learning systems.

As mentioned, this corpus is available through ELDA.

3.7. ERJ

ERJ (English read by Japanese) is a large corpus which contains utterances from Japanese speakers that read English text. This corpus was collected with the intention to support CALL research for language learning. Therefore the corpus provides elaborated pronunciation scores with the spoken utterances. The pronunciation of each student is rated regarding segmental, rhythmic and intonational aspects by native English language teachers.

The corpus is available at [37].

4. CLASSIFICATION OF DATABASES

In many cases potential users of these databases will have a clear understanding of what system they want and what they want to do with it. Systems can for example be speech recognizers, text to speech systems, pronunciation trainers or computer assisted language learning systems. The task might be to train, to adapt or only to test a system. Example applications are navigation devices, military communications, presentation systems or language learning systems. This sections tries to give additional help in finding suitable databases by discussing some areas of application and suggesting some databases for this task.

4.1. Speech operated Navigation Devices

Navigation devices, as most mobile devices still have to cope with limited computing power. Therefore systems running on these devices are less elaborated and usually only cover a restricted, command oriented user interface.

Of course, of major interest for navigation devices are city and street names as well as digits, for example for street numbers or postal addresses. Hence, a very interesting corpus for this task would be the CrossTowns corpus, as it covers mainly city names in a couple of languages. As it is not yet available, two further corpora can be recommended with restrictions: the CLIPS-IMAG and the ISLE corpus. The CLIPS-IMAG corpus has the advantage of covering the tourist domain, which is likely to contain similar places of interest as they will be demanded from navigation devices. The disadvantage of this corpus is, that it covers more or less exotic language combinations, that are unlikely to be in the focus of commercial products in the next years. Finally, the ISLE corpus. Compared to the other suggestions, it has the disadvantage not to contain in-domain data. Yet about half of the corpus are simple and short utterances, which is similar to the simple command interaction current navigation systems can handle.

4.2. Military Communications

This application area has the advantage that recently a couple of interesting corpora became available (see Section 3.3). The M-ATC (Military Air Traffic Control) covers pilot controller communications with a variety of accents, strong background noise and a high number of different speakers. The N4 corpus contains recordings from naval communication training sessions in the Netherlands. The transcriptions of the N4 corpus are very rich regarding information about speaker background. The Hiwire corpus finally contains spoken pilot orders that are input for the Controller Pilot Data Link Communications [38]. An advantage of this corpus compared to the two other ones is that the recordings were originally made in a studio. Thus this corpus provides clean speech as well noisy speech which was obtained through convolution of clean speech and noise. The HIWIRE and M-ATC corpus have the additional advantage to be free of charge for European researchers.

4.3. Speech operated Presentation Transcription Systems

There are two databases that are likely to be useful for this application, namely TC-STAR and TED. The TC-STAR corpus contains about 100 hours of Spanish and English transcribed parliament speech each. As listed in Table 2, this reduces to 11 hours of non-native English and some amount of non-native Spanish in both training and test corpora of TC-STAR. A larger part of the TC-STAR corpus is from non-native interpreters. As it is not clear to what extent speech from an interpreter relates to standard non-native speech the non-native interpreter part is not included in Table 2. The Translanguage English Database is a corpus which contains almost all presentations from the Eurospeech 1993. The speech material totals 47 hours, however only about 10 hours are transcribed. Due to the typical mixture of presentations from a variety of countries, it is believed that a large amount of the presentations is given with non-native accents.

4.4. Computer Assisted Language Learning Systems

Most speech technologies only need orthographic transcriptions for the databases to train systems. This is different for CALL systems. In order to detect and/or classify mispronunciation it is useful to have judgments of pronunciation quality

and/or a transcription at the phonetic level. Corpora which can provide proficiency ratings are the ISLE, Cross Towns, ATR-Gruhn, ERJ, Tokyo-Kikuko and CLSU corpus. Of these corpora, the ISLE and Cross Towns corpus contain also transcriptions at the phonetic level.

5. EXPERIENCES FROM DATA COLLECTIONS

The PhD thesis from Tomokiyo [2] and a paper from Schaden [39] are two examples of publications that give support on the collection of non-native databases based on their own experience. For detailed information, we refer to these works, this section summarizes some key findings.

Both publications report about anxieties of test speakers when being asked for speech recordings. Less proficient speakers are more afraid of recordings of their speech as they regard the situation as a test of their proficiency of the foreign language. Tomokiyo also reports that this becomes worse the more spontaneous the task is that the speakers have to perform. With prompts, the speech will keep its acoustic differences from native speech, such recordings however can not represent the different (wrong) grammar and word combinations non-native speakers typically produce. The fear reported above also occurs when texts for reading become too complicated. Special recording environments like a acoustic chamber further increase anxieties. A recording instructor sitting directly next or opposite of the speaker can ease the tension just by nodding after each utterance.

Further conclusions that Tomokiyo made are that the collector should be aware of the limited amount of speakers available. This effect is boosted by the fact that the speakers abilities to perform certain tasks vary. Thus the amount of suited speakers is reduced further. In many cases it might not be clear to data collectors what tasks will be regarded as easy or difficult by the foreign speakers.

For further research in this area it should also be clear that assessing the level of proficiency of the speakers is a very useful information. For example it might be possible (or necessary) to adapt systems to certain proficiencies of speakers, rather than creating one adapted system for speakers of one language speaking another language. Both, Tomokiyo and Schaden, came to this conclusion during their data collection.

6. CONCLUSIONS

The aims of this paper were to alleviate three issues of researchers when working on non-native speech. First, this paper provides an overview of a variety of existing databases. This should facilitate the search for appropriate databases for new research projects with non-native speech. Second, the databases described here cover a large amount of the databases that are used for publications in the area of non-native speech. Through the rather objective summarization of key aspects this helps to classify how significant the results of a certain paper are. Third and last, this paper has summarized some of the experiences from previous data collections of non-native speech that should help for future collections of non-native speech databases.

7. REFERENCES

[1] S. Witt, *Use of Speech Recognition in Computer-Assisted Language Learning*, Ph.D. thesis, Cambridge University Engineering Department, UK, 1999.

[2] L. Tomokiyo, *Recognizing Non-native Speech: Characterizing and Adapting to Non-native Usage in Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, Pennsylvania, 2001.

[3] K. Livescu, "Analysis and modeling of non-native speech for automatic speech recognition," M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999.

[4] "`http://en.wikipedia.org/wiki/Non-native_speech_databases`," 2007.

[5] J.C. Segura et al., "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication," 2007, `http://www.hiwire.org/`.

[6] T. P. Tan and L. Besacier, "A French non-native corpus for automatic speech recognition," in *LREC*, Genoa, Italy, 2006.

[7] R. Gruhn, T. Cincarek, and S. Nakamura, "A multi-accent non-native English database," in *ASJ*, 2004.

[8] University Munich, "Bavarian archive for speech signals strange corpus," `http://www.phonetik.uni-muenchen.de/Bas/`.

[9] D. Jurafsky et al., "The Berkeley restaurant project," in *Proc. ICSLP*, 1994.

[10] H. Ye and S. Young, "Improving the speech recognition performance of beginners in spoken conversational interaction for language learning," in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[11] T. Lander, "CSLU: Foreign accented English release 1.2," Tech. Rep., LDC, Philadelphia, Pennsylvania, 2007.

[12] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proc. ICASSP*, 2003.

[13] S. Schaden, *Regelbasierte Modellierung fremdsprachlich akzentbehafteter Aussprachevarianten*, Ph.D. thesis, University Duisburg-Essen, 2006.

[14] L. M. Arslan and J. H. Hansen, "Frequency characteristics of foreign accented speech," in *Proc. of ICASSP*, Munich, Germany, 1997, pp. 1123–1126.

[15] N. Minematsu et al., "Development of English speech database read by Japanese to support CALL research," in *ICA*, Kyoto, Japan, 2004, pp. 577–560.

[16] S. Fitt, "The pronunciation of unfamiliar native and non-native town names," in *Proc. of Eurospeech*, 1995, pp. 2227–2230.

[17] G. Stemmer, E. Noeth, and H. Niemann, "Acoustic modeling of foreign words in a German speech recognition system," in *Proc. Eurospeech*, P. Dalsgaard, B. Lindberg, and H. Benner, Eds., 2001, vol. 4, pp. 2745–2748.

[18] W. Byrne, E. Knodt, S. Khudanpur, and J. Bernstein, "Is automatic speech recognition ready for non-native speech? A data-collection effort and initial experiments in modeling conversational Hispanic English," in *STiLL*, Marholmen, Sweden, 1998, pp. 37–40.

[19] V. Fischer, E. Janke, and S. Kunzmann, "Recent progress in the decoding of non-native speech with multilingual acoustic models," in *Proc. of Eurospeech*, 2003, pp. 3105–3108.

[20] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in *LREC*, Athens, Greece, 2000, pp. 957–963.

[21] TNO Human Factors Research Institute, "Mist multilingual interoperability in speech technology database," Tech. Rep., ELRA, Paris, France, 2007, ELRA Catalog Reference S0238.

[22] S. Pigeon, W. Shen, and D. van Leeuwen, "Design and characterization of the non-native military air traffic communications database," in *ICSLP*, Antwerp, Belgium, 2007.

[23] L. Benarousse et al., "The NATO native and non-native (n4) speech corpus," in *Proc. of the MIST workshop (ESCA-NATO)*, Leusden, Sep 1999.

[24] Onomastica Consortium, "The ONOMASTICA interlanguage pronunciation lexicon," in *Proc. Eurospeech*, Madrid, Spain, 1995, pp. 829–832.

[25] C. Hacker, T. Cincarek, A. Maier, A. Hessler, and E. Noeth, "Boosting of prosodic and pronunciation features to detect mispronunciations of non-native children," in *Proc. of ICASSP*, Honolulu, Hawai, 2007, pp. 197–200.

[26] C. Teixeira, I. Trancoso, and A. Serralheiro, "Recognition of non-native accents," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 2375–2378.

[27] H. Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa, "TC-STAR: New language resources for ASR and SLT purposes," in *LREC*, Genoa, 2006, pp. 2570–2573.

[28] L.F. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillmann, "The translanguage English database TED," in *ICSLP*, Yokohama, Japan, Sep 1994.

[29] N. Mote, L. Johnson, A. Sethy, J. Silva, and S. Narayanan, "Tactical language detection and modeling of learner speech errors: The case of Arabic tactical language training for American English speakers," in *Proc. of InSTIL*, June 2004.

[30] N. Kikuko, "Development of Japanese speech database read by non-native speakers for constructing CALL system," in *ICA*, Kyoto, Japan, 2004, pp. 561–564.

[31] University Munich, "The Verbmobil project," `http://www.phonetik.uni\-muenchen.de/Forschung/Verbmobil/VerbOverview.html`.

[32] I. Trancoso, C. Viana, I. Mascarenhas, and C. Teixeira, "On deriving rules for nativised pronunciation in navigation queries," in *Proc. Eurospeech*, 1999.

[33] A. LaRocca and R. Chouairi, "West point Arabic speech corpus," Tech. Rep., LDC, Philadelphia, Pennsylvania, 2002.

[34] A. LaRocca and C. Tomei, "West point Russian speech corpus," Tech. Rep., LDC, Philadelphia, Pennsylvania, 2003.

[35] J. Morgan, "West point heroico Spanish speech," Tech. Rep., LDC, Philadelphia, Pennsylvania, 2006.

[36] I. Amdal, F. Korkmazskiy, and A. C. Surendran, "Joint pronunciation modelling of non-native speakers using data-driven methods," in *ICSLP*, Beijing, China, 2000, pp. 622–625.

[37] Speech Resources Consortium, "UME-ERJ English speech database read by Japanese students," `http://research.nii.ac.jp/src/eng/list/index.html`.

[38] Federal Aviation Administration, "Controller pilot datalink communications (CPDLC)," `http://tf.tc.faa.gov/capabilities/cpdlc.htm`.

[39] S. Schaden, "Casselberveetovallarga and other unpronounceable places: The CrossTowns corpus," in *Proc. LREC*, Genova, Italy, 2006.