

Chapter 2 – MILITARY SPEECH DATABASES

2.1 INTRODUCTION

Over the years, pronunciation variation due to non-native speech has been the interest of many phoneticians. A most remarkable number of researchers have studied the production and perception of the famous ‘l-r confusion’ for Chinese and Japanese natives. It is not an exaggeration to claim¹ that more than 50% of the research papers on non-native speech deal with this interesting subject of /l/ and /r/.

Long after phoneticians were drawn by the subject, non-native speech slowly started to become an issue in speech technology. Since speech databases are an invaluable resource for researchers in the field of speech technology, soon the first non-native speech databases were recorded. One of the problems with non-native speech is that there are potentially so many different kinds: if N is the number of languages in the world², the number of non-native accents is close to N^2 . This number only considers speech production. If the perception of speech is taken into account too, the number of possible language combinations scale with N^3 .

It is clear that in a field of research that has only recently started and with so many possible language combinations, the coverage by speech databases is rather limited. Yet, there are a number of interesting recordings available.

2.2 TERMINOLOGY

2.2.1 Language Specification

In non-native speech communication there are at least three languages of importance:

- 1) The native language of the speaker (S);
- 2) The language that is spoken, or the target language (T); and
- 3) The native language of the listener (L).

In literature one often finds the symbols L1 for native language and L2 for the target language, but this is not always used consistently and, especially in listening experiments, this terminology might lead to confusion. It is always a good exercise to really understand the configuration of languages in a description of non-native language experiments in a research paper. Van Wijngaarden [1] proposes the notation for communication between two persons:

$$S > (T) > L$$

Meaning that a speaker who’s native language is S speaks in language T to a listener whose native language is L . For the purpose of speech databases for speech technology, it usually suffices to specify S and T only, because L is either not considered or the technology takes the role of the listener, as is the case for speech, speaker, language and accent recognition.

¹ Based on JASA abstracts.

² N is estimated to be about 6000; the bible has been translated into 2000 languages alone.

2.2.2 Language Proficiency

One of the most important parameters in non-native speech communication is the language proficiency of the speaker and listener. There is a NATO standard, STANAG 6001, that classifies the language proficiency of people into five levels:

- 1) Elementary;
- 2) Fair (Limited working);
- 3) Good (Minimum professional);
- 4) Very good (Full professional); and
- 5) Excellent (Native/Bilingual).

These levels define both speaking and listening proficiency.

For speech databases, it is very important that the language proficiency of the individual speakers be known, because the quality and character of the speech is very dependent on this. None of the databases discussed in this chapter has classified the speakers according to STANAG 6001 levels, but there is generally information about the speaker's non-native language acquisition. Important information includes:

- **Native language:** The mother tongue of the speaker;
- **Age of acquisition:** The speaker's age when the non-native language was learned; and
- **Experience:** The number of years that the speaker has been regularly using the language.

Generally, an age of acquisition of over 6 years is considered to always lead to a noticeable non-native accent. All things being equal, the higher the age of first learning, the stronger the non-native accent will be. Of course, these parameters are not the only important factors in language proficiency; there are also matters such as willingness to learn another language, level of exposure, talent, etc. There are numerous cases where 'expatriates' live in a foreign country for decades without being exposed to the local native language at all.

Databases differ in the information that is specified about the speakers, even though there has been an effort to guide the database meta-data collection, such as through the EAGLES handbook [2].

2.3 A SELECTION OF AVAILABLE MILITARY SPEECH DATABASES

In this section an overview is given of some military databases that are available. There are many more recordings made of utterances under a wide range of conditions, for all the different studies made in literature. Here we only list the databases that have some relevance to speech technology research in military battlefield conditions.

2.3.1 FELIN Database

2.3.1.1 Overview

FELIN is a French database that was recorded in order to evaluate:

- How French infantrymen use speech-based command and control systems; and
- The performance of commercial systems on such a task.

The recording campaign lasted 3 days, involved 6 infantrymen in 11 exercises. The scenario was urban reconnaissance.

Each infantryman had his own recording device. Each device had 2 microphones: a bone-conducting microphone located at the top of the head inside the helmet and a boom-mounted microphone fixed on the helmet close to the cheek. One of the devices used was a speech command system. It recognized commands and replied using speech synthesis when needed. For the other devices, the speech recognition was done *a posteriori*.

All speech commands were preceded by the keyword “vocal” to initiate a dialog with the speech command system. The speech commands are divided in two types: those available for the whole group and those available for the group leader.

2.3.1.2 Technical Specifications

The database is made up of 15 hours 20 minutes of audio recordings.

All recordings have been transcribed with the annotation tool, Transcriber³, and saved in TRS format. All non-speech events are annotated.

Audio format: WAV – 16 kHz – 16 bits

Channel 0: bone-conducting microphone

Channel 1: boom-mounted microphone

2.3.1.3 Limitations of Use

The database distribution and usage is provided to the following restrictions:

- The database can only be used for academic, research and evaluation purposes;
- The database or the results of its use, cannot be used for any commercial purpose;
- The database cannot be redistributed without authorization of the database owner⁴;
- Any publication or evaluation result concerning the database must be communicated to the database owner;
- The anonymity of the recorded people must be guaranteed in any publication relative to the database; and
- The database must be destroyed upon request of its owner.

2.3.2 Canadian Soldier System Database

2.3.2.1 Overview

Experiments were carried out at Defence Research and Development Canada (DRDC) looking at communications among ground forces in an urban warfare environment as part of a future soldier system.

³ <http://trans.sourceforge.net/>.

⁴ Owner = DGA.

MILITARY SPEECH DATABASES

Audio recordings were made of the communications during these exercises. The exercises were done at two locations: in the laboratory using a computer gaming simulation and live exercises at a military facility.

The data was segmented into files with individual transmissions. All the files are labelled by time and speaker identification. The speaker id can be a number or a role, for example “2IC” for second in command. The data is divided into two directories for the laboratory simulation exercises and two directories for the field exercises. Each directory has up to 65 session directories with a number of individual transmission files in each. Some of the session directories contain log files with information about each transmission, but since the same information is contained in the file name there is no new information in the file.

The speakers are predominantly males; however, there are a few females in the dataset. There are 35 unique speakers in the field sessions and 7 speakers in the laboratory sessions. There is probably an overlap between the two groups, but this is unknown.

The laboratory data was recorded using head-set mounted microphones. Collection of the field data used a head-set microphone attached to a personal communications set. The recordings were done at a base station.

2.3.2.2 Technical Specifications

The data files are in PC WAV format with the following specifications:

- Encoding: linear PCM
- Sample rate: 44.1 kHz
- Resolution: 16 bits / sample
- Channels: 1

In order to maintain the time sequence of the files and to indicate the speaker, this information was included in the filename. The following filename format was used for the .WAV files for the laboratory sessions:

END TIME + PLAYER NAME + <space> + CHANNEL # + “.WAV”

END TIME = 6 digit time code when transmission ended, hhmmss

PLAYER NAME = “SC”, “2IC”, or “Player 2” to “Player 7”

CHANNEL # = 1 digit code

The filename format for the field experiments is similar and includes:

END TIME = 6 digit time code when transmission ended, hhhmmssss

PLAYER NAME = P##G# where G is the group number and P is the person number.

2.3.2.3 Limitations of Use

There are no restrictions on the distribution of this database; however, the database can only be used for research purposes. The producers of the data have asked that the National Research Council, Institute for Aerospace Research co-ordinate its distribution.

2.3.3 Dismounted Close Combat Database (DCCD)

2.3.3.1 Overview

The DCCD database was recorded between July and September 2001 by Aurix Ltd. (then called 20/20 Speech Ltd) for an UK MoD project for research into speech interfaces for dismounted infantry.

The trial data contains 16 male speakers who were soldiers in the British Army at the time of the recordings. The recordings consist of read speech (UK-English) consisting of alpha-numeric strings (using UK generic “number plate” format) and command-and-control type phrases. The speech was recorded during periods of external noise (e.g., gunfire) and of physical stress and (e.g., jogging). There are 11 distinct “acoustic” regions:

- 01) No additional noise; no physical stress.
- 02) Reverberation (Butts); no physical stress.
- 03) Gunfire and reverberation (Butts); no physical stress.
- 04) Gunfire (Firing-line); no physical stress.
- 05) Interfering speech; no physical stress.
- 06) Engine noise; no physical stress.
- 07) No additional noise; crawling and speaking.
- 08) No additional noise; jogging and speaking.
- 09) No additional noise; recovering and speaking.
- 10) No additional noise; speaking quietly.
- 11) No additional noise; shouting.

2.3.3.2 Technical Specifications

There are 47 speech files (dcc016-dcc063). Each speech file was recorded from a UK Infantry standard microphone as used for the personal role radio (this consists of a good quality noise-cancelling boom microphone that can be adjusted close to the speaker’s lips). The recordings were made on a DAT recorder at 44.1 kHz and later down-sampled to 20.50 kHz 16 bits PCM single channel.

The material spoken consisted of two types:

- 1) **Sighting Reports** consisting of reading between 1 and 6 “targets” each represented by the form of a string “*alpha one two three bravo charlie delta*” with report start and end notation. The following example is for 2 targets:
 - “target report begins, two targets sighted”
 - “charlie nine seven three juliet oscar bravo”
 - “juliet nine seven one sierra quebec uniform”
 - “report ends”
- 2) **Control Words** consisting of 30 phrases, each phrase consisting of between one and three words. Some examples are:
 - “left” “image” “toggle” “select” “mode” “transmit image” “toggle map image”

MILITARY SPEECH DATABASES

The speech files are currently recorded in a raw format and transcribed with the standard Transcriber tool. They occupy about 10 CDs. There is around 5.5 hours of speech comprising around 39,000 words in 10,575 phrases.

2.3.3.3 Limitations of Use

The database has not been used or released outside of Aurix Ltd. If this was required, it would need some additional work to formalize the documentation and double-check the directory structures.

2.3.4 Non-Native Military Air Traffic Control (nnMATC) Database

2.3.4.1 Overview

The Non-Native Military Air Traffic Communications (nnMATC) database was collected in Spring 2005 in the framework of the NATO/IST-013/RTG-031 task group, to support ongoing research in the field of speech processing under realistic battlefield conditions. Among those conditions, speakers non-nativeness and channel noise most heavily affect speech recognition performance. The nnMATC database combines the adverse effects of non-native speech and noisy environment, through realistic air traffic control communications recorded from an operational military Air Traffic Control (ATC) center. The nnMATC has been recorded at the time when it wasn't clear yet if we would succeed in releasing the Civilian ATC database (nnCATC) to the group. The nnMATC offers various advantages over the nnCATC, in particular it is a real military environment.

2.3.4.2 Characteristics

The nnMATC database consists of 24+ hours of contiguous ATC communications. These recordings were taped from the ATC center, implying a different speech quality depending on the speaker's location: on the controller-side, speech is mostly clean; on the pilot side, recordings suffer from a combination of background noise (cockpit) and communication interferences.

The non-native English accents covered on the controller side are mainly Belgian Dutch and Belgian French. On the pilot side, the variety is much wider with – among others – Dutch, Belgian Dutch, French, Belgian French, German, Italian, and Spanish accents. A few native American, British and Canadian English speakers are represented among the pilots as well. Most speakers are males, although a few female speakers are present as well, mainly among the controllers.

The nnMATC database was acquired through 13 sessions, all taking place at the same location but on different days. Multi-channel sessions – there are 12 of them – consist of multiple channels that have been recorded simultaneously for a few consecutive hours (typically 3-5 hours). Depending on air traffic conditions, 5 to 9 of those channels were active. The 13th session differs from the others as it relates to the recording of a single channel across a much longer period of time (a few days).

Most of these recordings originally suffered from long periods of inactivity – often up to several minutes. Therefore, silences have been trimmed down to 2 seconds. This process ended up in squeezing approximately 700 hours of raw audio material into 24+ hours of gapless speech (total time across all sessions and channels).

2.3.4.3 Technical Specifications

Audio bandwidth: 300 Hz – 3400 Hz (tapped over phone lines)

Recording format: wav, 22.05 kHz, 16-bit linear
Silence trimming: Signal < -40dbFS with a 1s post- and pre-roll margin
Total recording time: 24:34:04 (hh:mm:ss)
File format: nnMATC_sessionID_frequency.WAV (multi-session)
nnMATC_frequency_fileID.WAV (single session)

2.3.4.4 Limitations of Use

The nnMATC database is NATO UNCLASSIFIED. However, limitations of use apply. The nnMATC database is primary intended for research purposes within the NATO/IST-013/RTG-031 Group. Commercial use of the database is strictly prohibited. Identities involved in the recordings should be kept anonymous in any unclassified publication.

2.3.5 Non-Native Civilian Air Traffic Control (nnCATC) Database

2.3.5.1 Overview

The Non-Native Civilian Air Traffic Communications (nnCATC) was meant to be the reference ATC database for the NATO/IST-013/RTG-031 task group. As civilian data was supposed to be less sensitive to usage restrictions than its military counterpart, there was no plan to collect our own (military) database at the beginning of the project, but rather to collaborate with a European civilian air traffic control center to get access to ATC communications. Unfortunately, legal issues were more sensitive than expected and after one year of negotiations, no confidence was given that the data could ever be released. At that time, the group started to record its own military database (please refer to the nnMATC, described earlier). Eventually, an agreement was made and the civilian ATC database was released to the group under severe restrictions of use, coincidentally with the nnMATC.

Because of those heavy restrictions, we discourage the use of the nnCATC and prefer the nnMATC. The nnCATC remains however available for those interested in additional civilian ATC data.

Compared to the nnMATC, the nnCATC database offers a wider variety in non-native accents and preserves the absolute timeline (in other words, pauses between communications have not been removed). The amount of actual speech data is however much smaller, approximately 8 hours versus 24h.

2.3.5.2 Characteristics

The nnCATC was recorded on 3 radio frequencies (delivery, ground and departure) which are supposed to intercept the same flights. Each frequency was recorded for 20 hours continuously, ending up with 60 hours in total split into 30-minute segments. Each segment holds approximately 4 minutes of speech, on average.

Sound files were transferred from a digital archival system and suffer from audible aliasing (8 kHz – 16 bit). The full database is 3.22 GBytes in size.

2.3.5.3 Technical Specifications

Audio bandwidth: 300 Hz – 3400 Hz (phone line quality)

MILITARY SPEECH DATABASES

Recoding format:	wav, 8 kHz, 16-bit
Silence trimming:	None
Total recording time:	60:00 (hh:mm) – no silence trimming
File format:	chXX_msgYY.WAV with XX being the channel number (12, 20 or 46) and YY the message number (1 to 40)

2.3.5.4 Limitations of Use

Limitations of use apply, among others:

- Use of the database shall be restricted to research applications in the field of Speech and Language processing.
- Commercial use of the database shall strictly be prohibited.
- User commits itself to treat the information found in this database with confidentiality: the anonymity of the ATC data provider, the air traffic controllers, the pilots and the airline companies involved in the database shall at all times be guaranteed when disseminating the results of the scientific studies carried on the database.
- User shall not further distribute any of the contents of this database to any third party, without the prior written consent of Royal Military Academy (Belgium), the distributor of this database.
- User shall not publicly publish any part of the communications transcriptions in any form.
- User shall not use the contents of this database to take action against the RMA and/or the ATC Data Provider.

2.3.6 Destined Glory 04 Database (DG04DB)

2.3.6.1 Overview

Destined Glory 2004 (DG04) was a maritime expeditionary exercise conducted by STRIKEFORCESOUTH and includes live fire and a NATO Reaction Force Initial Operational Capability demonstration. The exercise was conducted on Capo Teulada range, an Italian Army armor training area, located on Sardinia's southern tip. The area is extremely remote. The exercise was conducted from 20 September to 16 October in 2004. Nearly 9,500 personnel, 50 ships and 46 aircraft participated. The Maritime and Amphibious forces were from 11 NATO nations.

The database created as a result of participation in the exercise has provided an excellent opportunity to assess next generation speech technology. During the exercise over 100 hours of raw audio were recorded. The exercise consisted of sea, land and air units that were available throughout the exercise. As this was an actual military exercise the primary communication devices were military push-to-talk radios. This imparts a real-world character to the communications, namely:

- Purposeful military verbal exchanges;
- Communication fading and multi-path;
- Communications contaminated with noise and interference;

- One-side of the communication; and
- Military vehicle artifacts.

Along with the environmental effects the NATO exercise allowed for a unique opportunity to capture specific human effects. These effects included personnel under physical stress, non-native speakers of English, and mid-communication language/word switching.

Although English and French are the official languages of NATO many of the communications were in the native language of the exercise participants. As such, not only are there examples of non-native English, but also a fair amount of Spanish, Italian, and Greek.

2.3.6.2 Technical Specifications

The data was accessed utilizing 4 Watkins Johnson (WJ-8611) receivers connected to a high quality 24 channel Mark Of The Unicorn 24I/O analog-to-digital converter connected to PCI-424. All audio files were recorded with 16 bits of fidelity at 48kHz sampling rate in little-endian PCM format. A portion of the data has been transmission marked, with speaker/call-sign markings, and transcribed. This has been accomplished with the Transcriber tool set. To date approximately 2 hours of raw audio has been processed. This contains 53 minutes of packed native/non-native English. Within the database are 57 speakers in 1176 transmissions. These transmissions contain 8277 total words of which 1042 are unique.

The audio file format can be decoded in the following way:

T01_285_AD_0822.pcm

T01 – Identifies the file as a DG04 audio file

285 – Julian date

AD – Air activity frequency D (G indicates ground activity)

0822 – Time of day collection began

The transcription files corresponding to the above audio file are the following format:

T01_285_AD_0822.trs

2.3.6.3 Limitations of Use

The DG04DB is NATO RESTRICTED and as such can only be used for research purposes within the NATO/IST-013/RTG-031 Task Group. Use outside of the NATO speech research community is strictly prohibited.

2.3.7 KFOR Text Corpus

2.3.7.1 Overview

In the ZENON project [3] an information extraction approach is used for the (partial) content analysis of written English HUMINT reports from the KFOR (Kosovo Force) deployment of the German Federal Armed Forces. Starting point of this development were 4,498 military reports (mostly in English) from the deployment. From these reports 800 were manually annotated and form the *KFOR Corpus* [4]. This corpus is a specialized micro text corpus.

MILITARY SPEECH DATABASES

The KFOR corpus is used for the following purposes:

- 1) It represents the basis for the construction of the information extraction component of the ZENON prototype. The lexicon and the grammars are optimized towards the corpus.
- 2) The performance of the ZENON information extraction is quantitatively evaluated relative to the KFOR corpus.
- 3) The KFOR corpus can be used for other research objectives (e.g., complexity of nominal phrases, word sense disambiguation, machine learning of grammatical structures, etc.).

2.3.7.2 Technical Specifications

The used annotation tool is GATE (General Architecture for Text Engineering, <http://gate.ac.uk>). The corpus covers 886,000 tokens and contains the annotations in different layers. The following layers are available:

- Original markups: In this layer those parts of the message are annotated, which are already formatted (e.g., addressee, topic, source).
- Token: This layer contains the annotations, which are supplied by the Tokenizer and the Part-of-Speech Tagger.
- Gazetteer: In this layer those expressions are annotated, which were identified over lists of names (e.g., first names, city names).
- Sentence: These annotations refer to sentences and begin and end markers of comments.
- Named entities: This layer contains the following annotation types: City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time and Title.
- Verb Group: The verbal phrases are annotated.

The corpus is represented in:

- The GATE-specific serial format ('SerialDataStore');
- The GATE-specific XML format ('XML serialization format');
- The XCES stand-off annotation format; and
- The TIGER-XML format.

2.3.7.3 Limitations of Use

The KFOR Text Corpus is classified (VS-NfD) and is not freely available.

2.4 REFERENCES

- [1] Wijngaarden, S., Steeneken, H. and Houtgast, T. (2001). "Methods and Models for Quantitative Assessment of Speech Intelligibility in Cross-Language Communication", In Proc. of the Workshop on Multilingual Speech and Language Processing, Aalborg, Denmark. (Available as NATO Publication RTO-MP-066, April 2003).

- [2] Howell, P. (1997). Handbook of Standards and Resources for Spoken Language Systems, Chapter 9: "Assessment methodologies and experimental design", pp. 344-380, Mouton de Gruyter.
- [3] Hecking, M. (2006a). "Content Analysis of HUMINT Reports". In: Proc. of the 2006 Command and Control Research and Technology Symposium (CCRTS) "The State of the Art and the State of the Practice", June 20-22, 2006, San Diego, California.
- [4] Hecking, M. (2006b). "Das KFOR-Korpus als Ergebnis semantisch annotierter militärischer Meldungen". Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 124.

