Design and characterization of the Non-native Military Air Traffic Communications database (nnMATC)

Stephane Pigeon[1]*, Wade Shen*[2] *and David van Leeuwen*[3]

[1]Royal Military Academy, Brussels, Belgium
[2]MIT Lincoln Laboratory, Boston, USA
[3]TNO Human Factors, Soesterberg, The Netherlands

Abstract

This paper describes the speech database that has a central role in the Interspeech 2007 special session "Novel techniques for the NATO non-native Air Traffic Communications database". The rationale for recording and distributing this common research object is given, and details about the acquisition and annotation are given, as well as some statistics. Further, a summary is given of potential uses of the database, in terms of evaluation measures and protocols.

1. Introduction

After having studied various aspects of speech in noise [5], speech under stress [6], and non-native speech [1], the current NATO research task group on speech and language technology, IST031-RTG013, has been studying the effects of all of these factors on various speech technologies. In order to facilitate speech research in this area, the task group collected a database of military Air Traffic Communication (ATC) during Tactical Leadership Programme exercises. The data was recorded in Belgium and the participation of several NATO nations guaranteed a wide variety of non-native English speech. Because the communications were between aircraft and the air traffic control center where the recording took place, there is also a wide coverage of signal to noise ratios. Finally, because speakers were functioning professionally in a NATO exercise, normal operating stress levels of the speakers can be expected.

The combination of non-native and noisy speech will make this challenging speech material for speech technology applications. We have attempted to define tasks that may be feasible to perform for certain areas in speech technology, but expect that new approaches may be required to deal satisfactorily with the data. The special session is meant to discuss approaches taken by the speech research community to deal with these challenges, and this paper can serve as a reference to the data collection and description.

The paper is organized as follows. In Section 2 we describe the details of recording and annotation of the database, in Section 3 we describe some statistical characteristics. We follow with a collection of proposed experiments that the design of the database allows for, with some proposed evaluation measures.

2. Recording of the database

The database was recorded at the premises of a Belgian Military Air Traffic Control Center (ATCC) in a period covered by Tactical Leadership Programme exercices, but also some regular days of operations. In total, 13 sessions were recorded. Twelve of them were *multi-channel* sessions, where up to 10 communication channels were recorded simultaneously for a period of 3–5 hours. An additional session was recorded where a *single channel* was monitored for several days continuously. For the multi-channel recording we used a Roland VS-2480 multi-track digital recorder, which sampled audio data at 32 kHz in 16 bit dynamic resolution. The compressed proprietary format data was later extracted and re-sampled to 22.05 kHz 16-bit PCM data and stored as RIFF/WAV sound files. The single channel recording was made using a Marantz PMD671 digital CF recorder, at sampled at 44.1 kHz and MP2 compressed at 128 kbs (mono, meaning that we work at the equivalent rate of 256 kbs, a high quality, if one considers stereo recording as a reference). This recording was later uncompressed and re-sampled to 22.05 kHz 16-bit linear PCM for distribution. Because the recorded channels regularly showed little activity, silent periods were automatically removed during the recording based on an absolute recording level threshold of -40 dBFS, with one second margins on both sides of the communication. Thus, continuous communication segments are separated by two second silences in the recordings. Communication activity timings were not preserved during the automatic silence removal process, so that is hard to reconstruct the order of the recorded communications from the multi-track recording. The communications were recorded by wire-tapping into the telephone switch that is used at the ATCC to route all radio communications to the controllers. Recordings were made '2-wire,' so that the ATC and pilot share the same

Part	name	size (h)	N_w	N_s	N_c
Training	trn	18.36	81239	520	5457
Development	dev	1.25	5707	62	438
Evaluation	tst	0.80	3094	40	181

Table 1: Proposed split of usage for various technologies, with basic statistic about number of words N_w, speakers N_s and call-signs N_c.

channel. The channels recorded were all military frequencies used at the ATCC.

In total over 700 hours of radio channel communications were monitored, from which over 24 hours of communication activity were recorded. A subset of approximately 20 hours was manually transcribed at the word level and annotated with speaker/listener entity identification. Further, aircraft call-signs were identified and marked in the transcription. Because parts of the the recorded speech were difficult to transcribe or were simply unintelligible, a semi-automatic process of alignment and rejection was used to filter regions of poor transcription. This filtered and annotated subset of the recorded data was split into three parts to be used for development automatic speech processing technologies, according to Table 1.

We have received clearance to share the database with speech researchers under the following conditions:

- The data is NATO unclassified,
- The use of the data is restricted to language and speech research,
- The NATO RTO IST031-RTG013 task group is the primary user, other researchers must contact the first author of this paper,
- Commercial use of the data is prohibited,
- All real identities must be kept anonymous in any unclassified publication.

Since the announcements of this Interspeech Special Session we have received 12 request for the database from outside our task group.

3. Characteristics of the database

Some basic annotation statistics of the database have been included in Table 1, in terms of number of words N_w, number of speakers N_s (as indicated by the speaker entity identification), and number of marked call-signs N_c. The speakers entity does not necessarily denote a unique speaker identity, because these were not known. The speaker identity identification was determined from either the channel (in case of an air traffic controller) or the call-sign (in case of a Pilot speaking).

An interesting notion of this database is that of the 'listener entity identification' (LEI). This is a label for the person who appears to be addressed in the communication. This information is extracted from the contents of the message, such explicitly named call-signs or ATC functions, or the context of the conversation. It was not always possible to extract the LEI information, in about 23 % of the utterances the LEI is not defined.

The entire database consists of 9833 utterances (i.e., transmissions), having an average word length of about 9.2 words/utterance, and an average duration of 3.74 seconds/utterance. In total 1708 unique words were transcribed, including multi-letter abbreviations such as AMS and IFF. In the transcriptions these words have been as such (A._MṠ.), differently from acronimes such as AWACS which are pronounced as a single word. For each of the words in the transcriptions pronunciation dictionaries are provided derived from the open source CMU pronunciation dictionary `cmudict-0.6d`. Two special 'words' [UNK] (unintelligible) and [BEEP] (transmission key beep) have been used in the transcriptions.

Most of the speakers in the database speak English with a non-native accent. Identified accents include Dutch, Belgian Dutch, French, Belgian French, German, Italian, and Spanish. A few Native American, British and Canadian English speakers are represented among the pilots as well. At the ATC side, the accent are mainly Flemish Dutch and (Belgium) French. The majority of speakers are males, most females are found among the controllers. Although it was the original intention to provide annotation of the spoken accents, this requirement was too hard to be met by the (North American) annotators.

4. Proposed experiments using the database

The NATO research task group had specific experiments in mind when collecting and annotating this database. In this section we will discuss these, along with their proposed evaluation metrics. Originally, these experiments were formalized in an *evaluation plan* [4], although a formal evaluation has not been held. Rather, this plan functioned as a guide to researchers for working on a common task of their interest, such that a maximum overlap in understanding approaches and challenges could be obtained.

4.1. Speech to text

With the basic segmentation and transcription, the nnMATC database provides training and evaluation test material for Automatic Speech Recognition (ASR) experiments. The optional dictionaries can be used to estimate this task's vocabulary and to have a default dictionary. Researchers can optionally use the evaluation set's segmentation information or even vocabulary for various experiments. The Word Error Rate (WER) is the pri-

mary evaluation metric. The WER is defined as the proportion of words incorrectly recognized after minimum-error word alignment, as can be carried out by the NIST tool `sclite` [2]. An obvious analysis conditioning is the source of the transmission, being either Pilot (noisy speech) or ATC (clean speech).

4.2. Call-sign identification

The call-sign identification task can be seen as a word spotting task, with the additional challenge that no explicit knowledge is given about the actual words. Rather, the syntax of call-signs is given as examples in the training set. For Call-Sign Identification (CSI) task the goal for a given system is to detect each occurrence of a spoken call-sign in each transmission. The system outputs a list of call-signs spoken during each transmission. This list may be empty if no call-signs are spoken. The performance metrics for CSI are the measures found in document retrieval, the precision

$$P = \frac{N_{\mathrm{corr}}}{N_{\mathrm{hyp}}} \tag{1}$$

and recall

$$R = \frac{N_{\mathrm{corr}}}{N_{\mathrm{c}}}, \tag{2}$$

where N_{corr} is the number of correctly retrieved call-signs, and N_{hyp} the number of hypothesized call-signs. A commonly used operating point that balances between the two measures is the F-measure

$$F_1 = \frac{2P}{P+R}. \tag{3}$$

Alternatively, the fraction of missed call-signs $P_{\mathrm{miss}} = 1 - R$ and the false alarm rate $R_{\mathrm{FA}} = (N_{\mathrm{hyp}} - N_{\mathrm{corr}})/T$, where T is the total duration of the speech, can be used to provide measures that can be interpreted in the 'spoken term detection' framework.

4.3. Speaker and Listener Identity Clustering

The Speaker Entity Identification (SEI) tags annotated provides support for speaker segmentation and clustering experiments. Segmentation can optionally be left out, as it is conceivable that in a real application meta-data, such at the radio's EM field strength, can be used to find contiguous segments originating from the same source. In fact, this database's automatic recording protocol implicitly depends on such meta-data by utilizing a carrier squelch.

The more interesting task therefore is the one of *speaker clustering*, for this database formally SEI-clustering because no true speaker labels are known. Special challenges to be considered are the wide variation of channel quality, which will have an impact on present-day speaker diarization algorithms.

In the task of SEI-clustering, the absolute identification of speaker entities is not required, rather, the task is to produce a consistent naming of the SEI labels. Alignment of the reference and hypothesis labels is carried out in such way, that the SEI clustering error is minimum. The SEI clustering error is defined as

$$E_C = \frac{N_{\mathrm{error}}}{N_{\mathrm{seg}}}, \tag{4}$$

where N_{error} is the number of segments incorrectly hypothesized and N_{seg} is the number of segments in the speech signal. This measure E_C is the segment-weighted counterpart of the time-weighted Speaker Diarization Error utilized in NIST Rich Transcription evaluations [3].

The nnMATC annotation files support a similar experiment to SEI clustering, namely *listener entity clustering*, or LEI. Although the evaluation metric and tools will be very similar to the SEI counterpart, the technology for LEI-clustering will most likely be different. The listener is defined as the entity to which the utterance is addressed. Unlike for SEI, the statistics of the speech acoustics cannot directly help in the LEI task, and rather the linguistic content and the context of the speech segment should be used.

Experimental conditions for LEI-clustering can include a text-based approach, where the transcripts of the transmissions are assumed to be known. Of course, researchers are encouraged to develop algorithms that only use the speech signal, and either use automatic transcripts or the acoustics of contextual segments to hypothesize the listener.

The long-term goal of SEI-LEI clustering is to be able to build automatic conversation tracking systems for ATC-type communication. This type of communication may also be observed in other centrally managed communication infrastructure, such as emergency services. Automatic tracking of conversations may help in communication error detection or with the analysis of critical situations.

5. Conclusion

This paper describes the rationale for, and the recording and annotation of the NATO nnMATC speech database, that has been distributed free of charge to the research community for the Interspeech Special Session on this database. It further describes some basic database statistics, and lists a number of proposed experiments, ranging from automatic transcription to the automatic extraction of the listener of a radio communication. Given the varying noise levels in the radio recordings, and the non-native accents of most of the speakers in this database, we consider the tasks very challenging. The fact that the communications are recorded during a NATO collaboration exercise guarantees a very realistic usage of language and talking style. We hope that this database will lead

to new approaches and some very interesting speech research.

6. References

[1] Laurent Benarousse, Edouard Geoffrois, John Grieco, Robert Series, Herman Steeneken, Hans Stumpf, Carl Swail, and Dieter Thiel. the nato native and non-native (n4) speech corpus. In *Proc. RTO Workshop on Multi-lingual Speech and Language Processing*, 2001. Aalborg.

[2] Jonathan Fiscus. `sclite`. `http://www.nist.gov/speech/tools/index.htm`. Software package.

[3] Jonathan Fiscus. The rich transcription 2005 spring meeting recognition evaluation. In *Proc. MLMI*, Lecture Notes in Computer Science, 2005.

[4] NATO IST031/RTG013. Plan for evaluation of speech and language processing technology on the NATO IST031/RTG013 military air traffic communications data set. `http://speech.tm.tno.nl/nn-matc/NATOevalplan.pdf`, 2007.

[5] Herman J. M. Steeneken and Andrew Varga. Comparison of assessment methods for automatic speech recognition in noise conditions. In *Proc. Speech Processing in Adverse Conditions*, pages 73–76, Cannes-Mandelieu, 1992.

[6] Isabel Trancoso and Roger Moore, editors. *Speech under Stress*, Lisbon, 1995. ESCA - NATO.