

Plan for Evaluation of Speech and Language Processing Technology on the NATO IST031/RTG013 Military Air Traffic Control Data Set

Version 1.0

1.0 Introduction

The goal of this document is to define the evaluation tasks, performance measures, and test corpora to support the first evaluation of speech and language processing systems on the NATO IST031/RTG013 Non-Native Military Air Traffic Control (nnMATC) Corpus. This evaluation is open to all interested volunteers and will support five core tasks:

- Speech-to-text (i.e. word transcription)
- Call-sign identification
- Native/Non-Native and accent detection
- Speaker entity identification (by call-sign or cluster)
- Listener entity identification (by call-sign or cluster)

2.0 Background

The NATO Research and Technology Organization (RTO) established the IST031/RTG013 Task Group in 2001 to assess the performance of speech technology in battlefield conditions. Whereas the application of speech technologies such as speech-to-text, speaker recognition and language recognition to broadcast news and conversational telephone speech has been extensively studied for the past 15 years, less attention has been focused on battlefield speech, which is typically characterized by noisy acoustic conditions, noisy channels, stressed speakers, and short duration utterances. Because NATO operations are typically multi-national, NATO battlefield speech will often contain speech spoken in multiple dialects and languages by both native and non-native speakers.

The Task Group has sought to collect and distribute one or more speech corpora having all (or most) of the characteristics found on the battlefield. The first such corpus is a military air traffic control corpus (nnMATC) provided by Belgian Royal Military Academy. This 24-hour corpus contains pilot-to-controller and controller-to-pilot transmissions. Almost all of the speech is spoken in English by both native and non-native speakers. The Task Group has transcribed and annotated this corpus and has packaged it for distribution to speech and language research sites. The corpus has been partitioned into training, development and evaluation partitions. A detailed description of the corpus is provided below in Section 4. The Task Group is likely to collect, transcribe,

annotate, and distribute other corpora (either from the battlefield or from conditions similar to the battlefield) in the near future.

3.0 Overview of the Task Definitions and Metrics

Five core speech processing tasks have been defined for evaluating speech and language processing on the nnMATC corpus. While these tasks have been defined specifically with the nnMATC corpus in mind, it is hoped that they might also be applied to other similar corpora. In this section, the task definition and metrics are introduced. Details of the system output formats and the evaluation of system performance are provided in later sections.

For each of the five tasks, a system is provided with two inputs. The first input is a digitally sampled speech waveform file containing multiple speech transmissions. The second input is a human-produced reference segmentation file indicating the starting time and duration of each individual speech transmission. While we realize that most applications of interest would not have the benefit of a reference segmentation file, we believe that systems exploiting readily available side information (e.g. radio frequency energy measurements) could perform high-accuracy automatic segmentation. Giving systems access to the reference segmentation allows us to focus the evaluation on the most interesting speech and language tasks and simplifies our performance metrics.

Though some applications might require systems that process transmissions causally, in this evaluation, we allow systems to process transmissions within a speech file non-causally. A system may both “look ahead” to future transmissions and/or use history information while processing a current transmission. Sites may choose to perform the entity clustering tasks from the audio input or from reference transcription (omitting the supplied call-sign annotation).

In all cases, performance is computed by comparing system output to a human-produced reference annotation containing “ground-truth” information.

3.1 Speech-to-Text

The goal of the speech-to-text (STT) task is to produce a word transcription of each speech transmission. The system outputs the sequence of words spoken. Non-lexical speaker sounds (e.g. cough, breath, etc.) and non-speech sounds (e.g. noise, tones, etc.) may occur in the speech transmissions, but these sounds should not be scored for evaluation.

System performance will be computed by aligning the sequence of hypothesized words against the sequence of reference words using the alignment strategy found in `sclite` (WER). The WER is the sum of the deletion, insertion and substitution errors divided by the number of words in the reference annotation.

In addition to the aggregate word error rate for a given system, word error rates will be computed for pilots and controllers separately since these channels suffer from significantly different levels of distortion.

3.2 Call-Sign Identification

On air traffic control networks, there are typically multiple speakers and multiple listeners. A given transmission always has only a single speaker, but it may be directed at one or more listeners. Additionally, successive transmissions will likely be spoken by different speakers and are likely to be directed at different sets of listeners. Speakers often identify themselves through the use of a call-sign, i.e. a sequence of words used as a unique identity for the airplane, pilot or controller. Some examples of call-signs are “Air France one fifty one heavy” and “Brussels approach”. Additionally, speakers often use call-signs to identify the identity of the airplane or controller position to whom they are directing the transmission. A given transmission may contain zero, one or multiple call-signs.

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 structure of call-signs and the handling of partially spoken call-signs are discussed in detail in later sections of these guidelines.

The CSI task will be evaluated using precision/recall metrics which are standard for named-entity recognition tasks.

$$Precision = \frac{\# \text{ call signs correctly detected}}{\# \text{ call signs hypothesized}}$$

$$Recall = \frac{\# \text{ call signs correctly detected}}{\# \text{ true call signs}}$$

These metrics are used within the information retrieval community and are related to miss and false alarm measures (commonly used in detection tasks). A common operating, the F-measure, point that balances between these two measures will also be used:

$$F_1 = \frac{2(Precision * Recall)}{Precision + Recall}$$

As the F_1 measure assesses a single operating point, additional call-sign error rate, $P_{fa}/P_d/P_{miss}$ analysis may be done to characterize system performance across a range of potential operating conditions that may be more representative of real-world operation.

3.3 Native/Non-Native Detection (N3D)

In NATO operations soldiers may speak their native languages when communicating with forces of their own country but are likely to speak in English when communicating with forces of other countries. Most automatic speech processing systems perform better when they are informed of the language of the speech input, and many automatic speech processing systems perform better when they are told whether the speech being processed is native or non-native. Thus, it is valuable to be able to determine automatically the language and nativeness of a speech utterance.

In the nnMATC corpus, all speech is spoken in English. However, much of the English is spoken by non-native speakers. For this data set we define two accent detection/identification tasks:

1. The Native/Non-Native Detection (N3D) Task – determine the likelihood that the speaker of a transmission is a native speaker of English. The system outputs a score of arbitrary scale according to the convention that higher scores indicate greater likelihood that a transmission was spoken by a native speaker than lower scores.
2. The Accent Recognition (AR) Task – For a given transmission, determine the accent of the speaker from a closed set of possible accents. For each test message, a system will generate a series of scores for each possible accent.

System performance is computed through the use of detection error trade-off (DET) curves. Transmissions are sorted by N3D/AR score. Probability of miss and false alarm are computed and plotted for all possible decision thresholds normalizing for detection priors.

3.4 Entity Clustering Tasks (Speaker Entity/Listener Entity)

Each ATC transmission can be associated with a speaker and an intended target. The Entity Clustering (EC) task is to identify the speaker/listener of each transmission. The identity can be an arbitrary string, but the system must use the same arbitrary string for representing the speaker in all transmissions produced by the speaker, and it must use a different arbitrary string for each speaker/listener. Sites should use the training data to seed speaker clusters for processing dev and eval data. Eval and dev sets have been chosen so that speakers have a significant prior probability of occurring in the train data.

To measure performance, the system speaker/listener identities are compared with the reference identities and an optimal entity mapping of reference speaker/listener identities to system speaker/listener identities is performed in a way that minimizes the EC error rate. The EC error rate is the sum of the EC errors divided by the total number of transmissions. For details about this process, refer to the description of NIST's `md-eval.pl`.

3.4.1 Speaker Entity Identification

Each ATC transmission is spoken by a person representing a particular “entity”. Pilots represent particular aircraft platforms (often identified by a call-sign). Often, a single pilot speaks all transmissions emanating from his aircraft. Sometimes, more than one pilot will speak transmissions from the same aircraft. The Speaker Entity Identification (SEI) task is to identify the entity represented by the speaker of the transmission either by call sign or a unique cluster-id represents the speaker. For each transmission, the system outputs an indication of whether the transmission was spoken by a pilot or by a controller.

Two SEI error measures are computed:

- Total SEI error is computed as the sum of SEI errors divided by the total number of transmissions. An error occurs if there is a pilot/controller confusion (i.e. system indicates pilot, reference indicates controller, or vice versa) or a pilot/pilot confusion (i.e. both system and reference indicate pilot, but the call signs don't match).
- SEI pilot/controller confusion error is computed by summing the SEI errors after ignoring the call-sign designations in both the system output and the reference annotation.

3.4.2 Listener Entity Identification

Each ATC transmission is directed to a particular “entity”. Controllers direct transmissions to particular aircraft platforms (often identified by a call-sign). The Listener Entity Identification (LEI) task is to identify the entity to which a transmission is directed. The system outputs an indication of whether the transmission was directed to a pilot or by a controller. For pilot-directed transmissions, the system must also specify the call-sign of the aircraft.

Some transmissions may be directed at all aircraft on the network. In such special cases, the system should indicate that the transmission was directed to all pilots.

Two LEI error measures are computed:

- Total LEI error is computed as the sum of LEI errors divided by the total number of transmissions. An error occurs if there is a pilot/controller confusion (i.e. system indicates pilot, reference indicates controller, or vice versa) or a pilot/pilot confusion (i.e. both system and reference indicate pilot, but the call signs don't match).
- LEI pilot/controller confusion error is computed by summing the LEI errors after ignoring the call-sign designations in both the system output and the reference annotation.

4.0 nnMATC Corpus Description

The Non-Native Military Air Traffic Control corpus (nnMATC) was collected in the spring of 2005 to support ongoing research in the speech processing under battlefield conditions. This corpus was collected during Tactical Leadership Programme (TLP) training exercises over thirteen different recording sessions.

Each session consists of recordings from twelve audio channels for durations of 3-5 hours (including dead periods). During post-processing, activity regions were extracted, yielding a total of 24+ hours of audio. Characteristics of this data are shown in the table below:

nnMATC Technical Specifications

Audio Bandwidth:	300 Hz - 3400 Hz (tapped over phone-line)
Recoding 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)

This corpus has been divided into three sets of ~16 hours, ½ hour and ½ hour between train, development and evaluation sets.

5.0 Data Processing Rules

The nn-mATC corpus has been partitioned into three sets: **train**, **dev** and **eval**. Final system performance will be evaluated using the **eval** set. Systems may use the training data to build acoustic models for ASR, generate long-term speaker models for clustering or any other training activity (setting thresholds, finding cohorts, etc.).

The development set has been set aside for system tuning and experiment prior to processing evaluation data. No tuning may be done using the evaluation set. As stated previously, data may be processed in any order in multiple passes. Sites are required to use the segmentation included in this corpus (see README for more details) for submission of results. UEM files will be provided with the appropriate reference segmentation information.

Sites wishing to perform evaluations of automatic segmentation and speech activity detector are free use the provided UEM files to assess segmentation performance, but for other tasks, submissions must make use of reference segmentation. Although, we encourage sites to report results for the call-sign and entity clustering tasks using speech-only, ASR and reference transcription as input, any or all of the input sources may be used.

6.0 Evaluation of Human Baseline

In order to help establish a human baseline of performance for the tasks listed above, sites may submit results from human subjects following a similar protocol to the protocol described above. The same submission format and tasks will be available to human subjects for baseline analysis.

All annotations should be done using DGA’s transcriber tool. Empty transcriber files, with segment marks will be provided as part of the nnMATC corpus. Please refer to the included README file for details. There are two differences between the human baseline process and the protocol described above.

1. *Training* – Human annotators should be given a set of training examples before starting the annotation of test set data. A set of example is provided with this corpus for practice. Refer to the included README file for details.
2. *Time Limit* – Subjects will be asked to process the test data within a four hour time slot. Please inform human subjects that they are free to take rest breaks as needed.
3. *Call Sign Entities* – Human subjects should label speaker and listener entities with call signs in a normalized form so that scoring of entity clusters can be done automatically using the `cseval.pl` tool as supplied by NIST.

7.0 Evaluation of Results

We will provide tools to evaluate the performance of sites on each of the individual tasks listed above. Table 1 shows the format of output expected for each of the tasks and the tools that will be supplied.

Task	Output Type	Tools
ASR	CTM	<code>Sclite</code>
LEI/SEI	RTTM	<code>md-eval.pl</code>
Accent Recognition/N3D	Accent Score Format (described below)	<code>n3deval</code> (described below)
Call-sign ID	RTTM	<code>cseval</code>

Table 1: Output formats for various tasks

Words that constitute callsigns should be joined with “_” and labeled as LEXEMES with their STYPE field marked as “callsign” in the output RTTM file. Non callsign words should be labeled as LEXEME with the STYPE field marked as “LEX.” Systems may annotate the confidence field (CONF) for use by `cseval.pl` for evaluation of P_{fa} , P_{miss} , P_d .

For speaker and listener entity extraction tasks call sign or cluster IDs may be used as described in section 3.4. For non-ASR outputs, sites should submit score files that contain vector of output scores/hypotheses with one utterance per line as follows:

<Utterance-ID> <ACCENT> <SCORE>

Where <ACCENT> must be one of native, english/american, french, dutch, spanish, german, italian or other. Sites may report scores on a subset of these accents.

For example:

```
Utterance-ID native -1.0
Utterance-ID english/american 0.9
Utterance-ID dutch 0.2
Utterance-ID french 0.0
Utterance-ID spanish -2.0
Utterance-ID german -9.0
Utterance-ID italian -9.0
Utterance-ID other 0.0
```

Sites may optionally provide a calibrated log-likelihood ratio as the score for CLLR computation.

- `Utterance-ID` – should be specified as presented in the provided segmentation.
- Scores for individual accent detectors. These will be used to compute det plots and decision thresholds. `other` should be used to score unmodeled accents.

ASR output should be provided in CTM format for information on this data format, please refer to NIST's `sclite` documentation. A standard GLM that is supplied with the data will be used to normalize orthographic and numeric issues during scoring.