Design and Development of Speech Corpora for Air Traffic Control Training

Luboš Šmídl, Jan Švec, Daniel Tihelka, Jindřich Matoušek, Jan Romportl, Pavel Ircing
Department of Cybernetics, University of West Bohemia
Univerzitní 8, 306 14 Plzeň, Czech Republic
{smidl, honzas, dtihelka, jmatouse, rompi, ircing}@kky.zcu.cz

Abstract

The paper describes the process of creation of domain-specific speech corpora containing air traffic control (ATC) communication prompts. Since the ATC domain is highly specific both from the acoustic point-of-view (significant level of noise in the signal, non-native English accents of the speakers, non-standard pronunciation of some frequent words) and the lexical and syntactic perspective (prescribed structure of utterances, rather limited vocabulary), it is useful to collect and annotate data from this specific domain. Actually, the ultimate goal of the research effort of our team was to develop a voice dialogue system simulating the responses of the pilot that could be used for training aspiring air traffic controllers. In order to do so, we needed – among other modules – a domain-specific automatic speech recognition (ASR) and text-to-speech synthesis (TTS) engines. This paper concentrates on the details of the ASR and TTS corpora creation process but also overviews their usage in preparing practical applications and provides links to the distribution channel of the data.

Keywords: speech corpus, speech recognition, text-to-speech, air traffic control communication

1. Introduction

The air traffic control (ATC) constitutes a crucial segment of the whole air traffic industry – the air traffic controllers communicate with the pilots almost continuously in order to ensure the fluent and safe flow of the aerial traffic. The job of a controller is very demanding and requires – besides the specific personal prerequisites – an intensive training.

This training is mainly focused on teaching and reinforcing the communication skills of the aspiring controller. The current state-of-the-art training procedure (at least in the Czech Republic) involves so-called *pseudopilots*. These are usually retired pilots that prepare training scenarios and consequently act as pilots of virtual plane (usually more than one at a time), communicate with the controller in training (*trainee*) and process the spoken prompts received from trainees to the form that can be entered into the software that simulates the plane movement on the radar screen.

Two major drawbacks were identified in such a training setting. The first is rather obvious – the length of the controller training (approx. 2 years on average) and the relatively high salaries of the pseudopilots make the whole process very expensive. This was actually the first incentive that sparked the idea of developing an automatic training simulator based on the intelligent spoken dialogue system. Only after delving into the specifics of the ATC communication, we have realized that the scenario involving the pseudopilots in actually quite unrealistic in several aspects:

First, In the real ATC scenarios, the controller will need to understand the English utterances pronounced mostly by non-native speakers (remember that we are talking about the air space situated in the Central Europe), sometimes with quite an exotic accent. On the other hand, during the training sessions involving the human pseudopilots, it is usually the case that Czech trainees are listening to the English utterances pronounced by a retired Czech pilot and thus the mutual understanding is naturally much easier. Moreover, the training environment lacks the noise that is massively present in the real-world VHF radio communication; this might lead to a drastic decrease of the unprepared controller's ability to understand the communication once he is put into service.

On the other hand, the human pseudopilot usually handles several virtual airplanes. This might result into confusion of the trainee as he hears the same voice from different simulated aircrafts.

So, when designing the spoken dialogue system (let us call it the *artificial pseudopilot – aPP* and see its simplified block diagram on Figure 1) that should replace the human pseudopilot, we tried to rectify the shortcomings mentioned above.

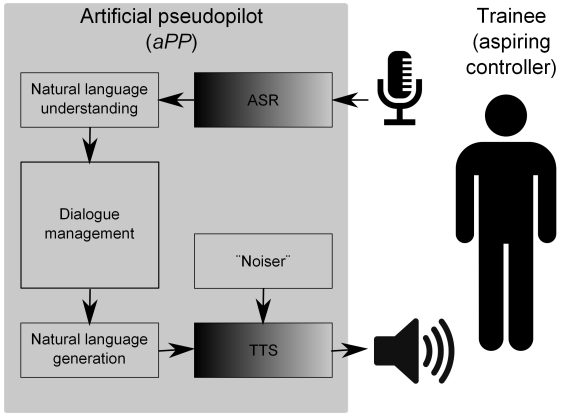

Figure 1: Block diagram of the artificial pseudopilot

All the specifics of the ATC communication that we have listed are actually connected with the "interface" blocks of the *aPP* – that is, the automatic speech recognition (ASR) and text-to-speech (TTS) modules. The state-of-the-art methods for development of those modules naturally require large speech corpora for either training the acoustic

and language models (ASR) or building the inventory of basic speech units (TTS). Since the ATC data are only rarely being collected[1], we have decided to design and collect our own corpora, taking into account the peculiarities of the ATC communication and tailored to our specific needs.

2. ASR Corpora

We took advantage of the fact that one of our industrial partners develops complex IT solutions for several ATC authorities and airports and, as such, has access to the ATC communication recordings. We were therefore able to secure almost 140 hours of recorded communication in the following structure:

- GRP (ground control) – communication before takeoff and after landing – 19.2 hours of data
- TWR (tower control) – communication during takeoff, landing and landing standby – 22.5 hours
- APP (approach control) – communication during landing approach – 25.5 hours
- ACC (area control) – communication during overflights and cruises – 71.3 hours

Those data were first segmented and the segments were classified as either speech or non-speech using an in-house voice activity detector (Prcín et al., 2002). The segments classified as speech were consequently imported into the Webtransc annotation tool (Valenta and Šmídl, 2015) that serves for online annotation of multimedia data. It enables annotators to play the segments, to transcribe their content (as well as to add various markers dedicated for non-speech events) and to add several types of metadata (such as the speaker's communication role – pilot or controller, in this case). The screenshot of Webtransc is shown on Figure 2.
The majority of the employed annotators already had some experience with speech corpora annotation. We have nevertheless prepared a detailed transcription manual, paying special attention to instructions that concern handling of non-standard pronunciations, special ATC terminology, spelling alphabet and other issues peculiar to ATC.
Here are the most interesting instructions from the manual:

- the utterances from pilots are marked as *Air* whereas the controllers' utterances are tagged with the label *Ground*. Naturally, this distinction is very important as both channels have significantly different acoustic qualities.
- the words pronounced in a non-standard way are manually equipped with the actual phonetic transcription (written using the Arpabet transcription code[2])
- the numerals pronounced correctly according to the ATC protocols[3] are written simply as numbers separated by space – otherwise they are also equipped with a actual phonetic transcription. The same notation is used for letter that are spelled out using the spelling alphabet. For example, the transcription `3 0 3 V direct to H D O` means that this particular fragment was pronounced "*tree zero tree victor* direct to *hotel delta oscar*"
- three broad classes of *non-speech events* were identified and marked in the transcriptions:
  - *hesitation sounds* made by the speaker – e.g. "ehm", "uh-huh"', "hmm", etc.
  - non-speech sounds produced by the speaker – e.g. coughing, lip-smacking, exceptionally loud breath, laughing, etc.
  - environmental noises that stand out from the (already quite high) natural noise in the background

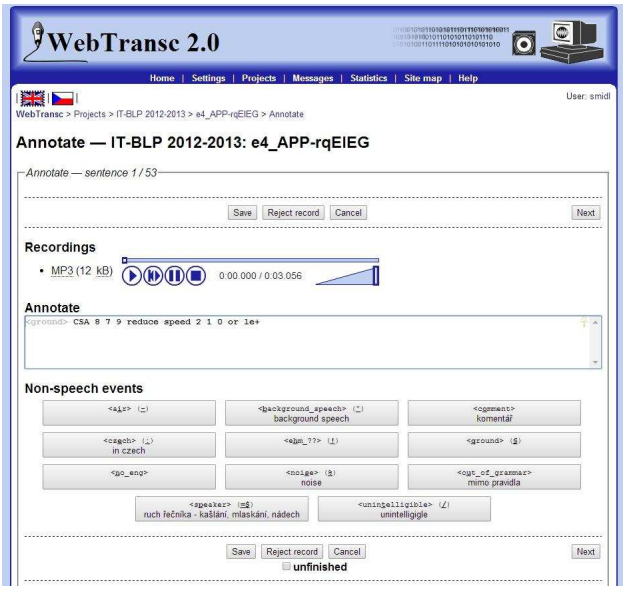

Figure 2: Screenshot of the WebTransc tool

After the transcription phase, the data samples were checked and, based upon this inspection, several rewriting rules (in the form $A \rightarrow B$) were designed to correct the most common typos (such as "aftrenoon" to "afternoon") and unify possibly inconsistent transcription of certain names (e.g. "Germanwings" to "German_Wings").
The unique words from such cleaned and normalized transcripts were then used as a basis for creation of the pronunciation lexicons. The actual pronunciations (phonetic baseforms) are either:

- extracted from the existing pronunciation lexicons available for English, namely
  - BEEP (BrE)
  - CALLHOME (AmE)
  - SWITCHBOARD (AmE)
  - CMU Dict (AmE)
- taken from transcripts where they could be written by the annotators

[1] The only other ATC corpus known to authors is the Air Traffic Control Complete (Godfrey, 1994) which is quite dated and also of rather poor technical quality.

[2] `http://en.wikipedia.org/wiki/Arpabet`

[3] See for example `http://aviationknowledge.wikidot.com/aviation:nato-phonetic-alphabet`

- manually created by an expert in cases when neither of the above sources provided any outcome

The resulting lexicon also employs the above mentioned Arpabet transcription code that is (most notably) used in the CMU Dict. There is a one-to-one mapping between Arpabet and the International Phonetic Alphabet (IPA) symbols. The portion of the speech recordings (approx. 20 hours) with corresponding transcriptions obtained in the way described in the previous paragraphs was released via the LINDAT/CLARIN repository (Šmídl, 2011). The pronunciation lexicon together with unigram, bigram and trigram word counts that can be directly used for language modeling was released later on via the same distribution channel (Šmídl, 2013).

3. TTS Corpora

For the reasons that were already mentioned in the introduction, in order to keep the artificial pseudopilots realistic, we should prepare a relatively broad portfolio of English TTS voices with various accents. Standard commercial high-quality unit-selection TTS voices are based on speech corpora counting more than 10 thousand phonetically rich sentences recorded by a professional speaker in a sound studio, often with an electroglottograph. Recording of such amounts of data is very time consuming (and thus costly) both for the speaker and for the subsequent expert annotation. Moreover, if the speaker is supposed to record in English which is not his/her native language, practical difficulties and expenses increase even more.

Therefore, we have decided to utilize a domain-specific approach to TTS voice creation, relying on the fact that the natural language generating module of the artificial pseudopilot generates utterances only from the restricted domain of typical ATC radio communication (albeit the number of potential different utterances is unlimited). The key to high quality TTS in such a scenario is to record a speech corpus that is both small in size and at the same time rich in variability – that is, it covers well the domain as the whole and/or allows high quality generation of the missing pieces. First step is to select chunks of texts (not necessarily the whole sentences) to be recorded. Using the algorithm described in (Jůzová and Tihelka, 2014), we have selected 1 000 chunks[4] from more than 34 000 transcripts of the speech uttered by pilots that appeared in the ASR corpus described in Section 2.. This limited set still covers 46.6% of types (unique words) of the original text data.

For the actual corpus recording, we have used the tool described in (Matoušek et al., 2008). Since the texts selected for recording are, due to the selection algorithm, actually just the fragments of sentences (chunks) that might not be always meaningful, it can be difficult for speaker to record them with a required prosody. In order to help the speaker to adhere to the correct prosody style during recording, we grouped the chunks derived from a particular phrase together, and presented them in relation with their source phrase. That way the speaker has the chance to read the phrase as a whole, and then to read the individual chunks in the same style as he/she read the chunk within the whole phrase.

The first *aPP* voice was recorded by a non-professional Taiwanese male speaker with a very typical and strong Mandarin accent in English. In order to test the "worst-case scenario", we deliberately did not use a professional sound studio but recorded the utterances in a standard office room instead, using a regular PC with a high-end external sound card and a microphone; no electroglottograph signal was recorded. The recording of the previously mentioned 1 000 text chunks took about 10 hours that were split into two days of recording. The annotations were subsequently manually checked and corrected where necessary. The pitch marks (i.e, the glottal closure instants) were identified directly from the speech signal using an algorithm described in (Legát et al., 2011).

The resulting corpus was used for the development of the actual TTS system (see Section 4. for details) and also released via the LINDAT/CLARIN repository (Jindřich Matoušek, 2014b). The released version includes – besides the obvious speech files and corresponding transcriptions – the information about the pitch marks, the pronunciation lexicon and the corresponding phonetic alphabet (the Arpabet transcription code, the same that was used in the ASR lexicon mentioned in Section 2.).

Later on, the the procedure described in this section was used to prepare and release the corpora containing the domain-specific English voices with Serbian (Jindřich Matoušek, 2014a), Czech (Jindřich Matoušek, 2015a) and German (Jindřich Matoušek, 2015b) accents.

4. Usage of the Corpora for Application Development

The ASR corpus described in Section 2. was used to train several versions of the automatic speech recognition system. All of them utilize the Hidden Markov Model (HMM) architecture for acoustic modeling and word n-grams as language models. The initial acoustic model was trained using only 17.5 hours of annotated speech data where the utterances of the pilots (denoted *Air*) and controllers (*Ground*) were mixed together. This baseline model was used essentially for "sanity checking" and setting up the basic acoustic model parameters (such as the number of HMM states and Gaussian mixtures). The next set of acoustic models consists of models trained separately for *Air* and *Ground* data, using the entire corpus of transribed speech (except for the 1.4 hours that were put aside as a test set). The ASR results achieved on the test set are summarized in Table 1.

Data source	Training data size [hours]	WER [%]
Air	54.9	25.27
Ground	78.8	7.59

Table 1: ASR results for individual data sources

The table shows that the recognition performance for controllers' utterances is far better than the the one for the pilots' data. This was of course to be expected as the data

[4]The number of selected chunks was chosen rather arbitrarily – the motivation was not to exceed two days of corpus recording.

are recorded at the control tower and thus the ground control speech is naturally acoustically much cleaner than the speech transmitted from the planes via radio. It was also good news for the prospective development of the *aPP* dialogue system as it is supposed to recognize controllers' speech, not pilots' (cf. Figure 1).

The acoustic models for the speech recognition module that is used in the actual dialogue system were trained using the ASR speech corpus described in this paper (140 hours in total) augmented with additional 460 hours of LibriSpeech data (Panayotov et al., 2015). The final set of models employs three-state HMMs with 2 000 states in total and 16 Gaussian mixtures per state. The demo of this ASR system can be found at `https://itblp.zcu.cz/asr-sc/`.

The TTS systems developed for the *aPP* dialogue system are based on the concatenative speech synthesis paradigm, employing unit-selection algorithms with diphones as basic units. In order to enhance the realistic feeling of the artificial pseudopilot's speech, a module capable of adding several types of noises typical for the ATC communication was also designed. This module can simulate:

- various background noises that last for longer periods of time and overlap with the speech
- short noises like the sound of the transceiver switching on/off, various on-board messages and beeps, etc.
- sudden or graduate changes in the signal volume and even the total signal outage where it replaces the signal with a noise

and many more.

The resulting synthetic speech produced by the TTS system build from the corpus recorded by the Taiwanese speaker was evaluated in several listening tests. Preliminary small-scale listening tests showed acceptable overall synthetic voice quality and very identifiable and realistic Mandarin accent. Especially when the synthetic speech is mixed with simulated typical radiocommunication and cockpit ambient noise, the result is very convincing. However, a relatively high number of problematic or even unintelligible synthetic speech segments (words) were observed. Therefore, more formal listening tests were carried out, with two independent listeners evaluating 500 testing sentences from the held-out data of the source sentence database. The listeners were instructed to note and localize all disturbing synthetic artifacts in the generated speech causing disfluencies, unintelligible segments or just subjectively uncomfortable phenomena. This procedure resulted into approx. 50 manual interventions into the speech corpus annotation and especially segmentation, showing that most of the problems were caused by local failures of the automatic phonetic segmentation algorithm.

After these interventions, the quality and intelligibility of the synthetic speech improved significantly, which means that our method of minimalistic domain-specific speech corpus recording of non-native English speakers seems to be practically useful for acquisition of a rich inventory of non-native synthetic pseudopilot voices in the ATC simulator. The demo of the TTS system (with actually more foreign accents of English than it is presented in this paper) can be found at `http://itblp.zcu.cz/tts/index.html`.

5. Conclusion

The paper presented the motivation and the process of creation of the speech corpora that can be used for developing ASR and TTS applications in the domain of air traffic control. The evaluation experiments of the ASR and TTS systems build upon the corpora described in this paper have shown that the presented language resource are indeed useful for practical application. Both system, evaluated within this paper as stand-alone systems, were also already incorporated into the full *aPP* dialogue system. Its description is outside the scope of this paper and can be found in (Šmídl et al., 2016) and (Stanislav et al., 2016). However, the interested reader can test it at `https://itblp.zcu.cz/app-demo`.

All the corpora described in the paper are available in the LINDAT/CLARIN repository (see Section 8. for individual links)

6. Acknowledgments

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic projects LINDAT/CLARIN (No. CZ.02.1.01/0.0/0.0/16_013/0001781) and PUNTIS (No. LO1506).

7. Bibliographical References

Jůzová, M. and Tihelka, D. (2014). Minimum text corpus selection for limited domain speech synthesis. In Petr Sojka, et al., editors, *Text, Speech and Dialogue*, volume 8655 of *Lecture Notes in Computer Science*, pages 398–407. Springer International Publishing.

Legát, M., Matoušek, J., and Tihelka, D. (2011). On the detection of pitch marks using a robust multi-phase algorithm. *Speech Communication*, 53(4):552–566.

Matoušek, J., Tihelka, D., and Romportl, J. (2008). Building of a speech corpus optimised for unit selection tts synthesis. In *LREC 2008, proceedings of 6th International Conference on Language Resources and Evaluation*, pages 1296–1299. ELRA.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*. IEEE.

Prcín, M., Müller, L., and Šmídl, L. (2002). Statistical based speech/non-speech detector with heuristic feature set. In *6th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2002) / 8th International Conference on Information Systems Analysis and Synthesis (ISAS 2002)*, pages 264–269, Orlando, FL.

Stanislav, P., Šmídl, L., and Švec, J. (2016). An automatic training tool for air traffic control training. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 782–783.

Valenta, T. and Šmídl, L. (2015). Webtransc — a WWW interface for speech corpora production and processing. In Andrey Ronzhin, et al., editors, *Speech and Computer: 17th International Conference, SPECOM 2015, Athens, Greece, September 20-24, 2015, Proceedings*, pages 487–494. Springer International Publishing, Cham.

Šmídl, L., Chýlek, A., and Švec, J. (2016). A multimodal dialogue system for air traffic control trainees based on discrete-event simulation. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 379–380.

8. Language Resource References

John J. Godfrey. (1994). *Air Traffic Control Complete*. Linguistic Data Consortium – LDC, 1.0, ISLRN 367-677-522-995-8.

Jindřich Matoušek, Daniel Tihelka. (2014a). *English TTS speech corpus of air traffic (pilot) messages - Serbian accent*. University of West Bohemia, distributed via LINDAT/CLARIN, IT-BLP resources, 1.0, ISLRN http://hdl.handle.net/11234/1-1462.

Jindřich Matoušek, Daniel Tihelka. (2014b). *English TTS speech corpus of air traffic (pilot) messages - Taiwanese accent*. University of West Bohemia, distributed via LINDAT/CLARIN, IT-BLP resources, 1.0, ISLRN http://hdl.handle.net/11234/1-1461.

Jindřich Matoušek, Daniel Tihelka. (2015a). *English TTS speech corpus of air traffic (pilot) messages - Czech accent*. University of West Bohemia, distributed via LINDAT/CLARIN, IT-BLP resources, 1.0, ISLRN http://hdl.handle.net/11234/1-1587.

Jindřich Matoušek, Daniel Tihelka. (2015b). *English TTS speech corpus of air traffic (pilot) messages - German accent*. University of West Bohemia, distributed via LINDAT/CLARIN, IT-BLP resources, 1.0, ISLRN http://hdl.handle.net/11234/1-1588.

Luboš Šmídl. (2011). *Air Traffic Control Communication*. University of West Bohemia, distributed via LINDAT/CLARIN, IT-BLP resources, 1.0, ISLRN http://hdl.handle.net/11858/00-097C-0000-0001-CCA1-0.

Luboš Šmídl. (2013). *ATCC: Pronunciation lexicon and n-gram counts for ASR module*. University of West Bohemia, distributed via LINDAT/CLARIN, IT-BLP resources, 1.0, ISLRN http://hdl.handle.net/11858/00-097C-0000-000D-EC92-F.