PATSY-I: A Corpus on Non-Native English Air Traffic Communication

Caroline Kaufhold[1], *Christine Martindale*[1], *Axel Horndasch*[1]
Klaus Reinhard[2], *Elmar Nöth*[1]

[1]Pattern Recognition Lab, Friedrich-Alexander University Erlangen-Nuremberg, Germany
[2]e.sigma Technology GmbH, Ilmenau, Germany
{caroline.kaufhold, christine.f.martindale, axel.horndasch, elmar.noeth}@fau.de,
kreinhard@esigma-technology.com

Abstract

In many global tasks English is used as an international language. As a consequence, non-native speakers of the English language often communicate with other non-native speakers. An example is the Air Traffic Control (ATC) service which directs aircrafts on the ground and through controlled airspace. It is of course essential that there is a perfect understanding between the pilot and the ground-based controller to prevent collisions and to organize air traffic efficiently. Aviation English already accommodates non-native speakers of English by providing guidelines for wording and phraseology. To avoid confusion, for example, letters and numbers are spelled according to the international spelling alphabet provided by the International Civil Aviation Organization (ICAO). However, the ability to speak and understand English still has a high impact on communication success.

In this paper, we present a corpus that was recorded in the context of the ATC phraseology training system PATSY, the prototype of which was presented at the Show&Tell session at Interspeech 2015 [1]. The corpus consists of basic ATC utterances by speakers of 16 different mother tongues. Furthermore, "Please Call Stella" [2] and part of "The Rainbow Passage" [3] were recorded twice for every speaker with different biases. We plan on using this data to study the entrainment effect, which was observed for conversations by Levitan and Hirschberg [4]. Preliminary results on basic ATC utterances show a moderate correlation between the speakers' self-assessment and the GoP (Goodness of Pronunciation) score.
Index Terms: English as Lingua Franca (ELF), Computer Assisted Pronunciation Training (CAPT), Accent Entrainment

1. Introduction

A multitude of circumstances arise in daily life where people with different mother tongues must communicate with each other. However, depending on the context, the consequences of these interactions vary greatly regarding severity [5]. This project investigates the use of English as an international language, or as the Lingua Franca (ELF), in the context of air traffic control, in which flawless communication is of course of utmost importance.

Communication between speakers where at least one party's mother tongue is not English can be classified into two distinct categories: firstly, one of the interlocutors is a native speaker (NS) and the other is a non-native speaker (NNS) and secondly, both interlocutors are non-native speakers. While miscommunication is likely if only one of the interlocutors is a NNS, communication success relies even more on the intelligibility of the parties' spoken English if both are non-native speakers.

The paper is structured as follows: section 2 gives a general introduction to the work-in-progress phraseology training system PATSY [1] which was initially presented at the Show&Tell session at Interspeech 2015. The recording setup, structure of the recordings and the collected data is then discussed in section 3. The part of the corpus which contains basic ATC wordings is described in more detail in section 3.2. The other part, which will be used for investigating the effect of accent entrainment is described in section 3.3. Our experimental setup and preliminary results for automatically assessing pronunciation proficiency on the PATSY-I corpus using the Goodness of Pronunciation (GoP) score [6] are discussed in section 4. Section 5 summarizes the PATSY-I corpus and concludes the paper.

2. PATSY Project

PATSY is an abbreviation for the German name of the project "Piloten/ATC Trainingssystem für den Sprechfunk" which translates to "Pilot/Air Traffic Controller (ATCO) Training System for radio communication". Air Traffic Control (ATC) is the service provided by the ground-based controller or ATCO, who navigates the aircraft on the ground and in the controlled airspace. Pilots and ATCO maintain radio contact and communicate in the English language. In order to compensate for difficulties in understanding due to the language barrier, the International Civil Aviation Organization (ICAO) introduced special spelling rules for letters and digits. Phraseology and the special pronunciation rules are taught in flight school and it is planned that PATSY will become part of the training for new pilots. On the one hand PATSY shall help the trainees to internalize the vocabulary and syntax used in ATC communication. On the other hand, it will give direct feedback regarding the user's pronunciation and intelligibility.

In our first prototype, which was shown on Interspeech 2015 [1], we use the Goodness of Pronunciation (GoP) score [6] to assess pronunciation skills. During the assessment, the user can listen to the play-back of a reference speaker after each turn. By re-recording the utterance he/she can improve his/her pronunciation. PATSY will then compute the new pronunciation score and update the evaluation presented to the user. Over the last year, we have continued to collect recordings from as many different mother tongues as possible. The "PATSY-I" corpus, which contains the data collected thus far, is discussed in the following section.

3. The PATSY-I Corpus

The PATSY-I corpus is the first set of recordings collected in the course of the PATSY project. The corpus recordings can be split into two parts: recordings of type "Flight Basics" and recordings of type "Read Paragraphs". Basic ATC recordings comprise ICAO spelling alphabet words and numbers with and without special pronunciation rules. They will be referred to as "Flight Basics". The other part of the corpus consists of recordings of the two paragraphs: "Please Call Stella" [2] and a part of "The Rainbow Passage". These recordings will be referred to as "Read Paragraphs" [3].

Our goal was to gather recordings from speakers with many different mother tongues with a reasonably high level of proficiency in English. Therefore, many of the participants were PhD students working at the Pattern Recognition Laboratory of the Friedrich-Alexander University Erlangen-Nuremberg. Recordings of 70 speakers were collected of whom 47 were working in research and 23 were students or graduates at the time of the recordings. As shown in in Table 1, there are 34 German and 17 Chinese speakers among the 70 participants. The remaining 19 have a wide range of mother tongues represented by three or less speakers. 54 of the 70 speakers are male and 16 are female. The recording setup and a description of the reading tasks is given in the following section.

$\sum$	**de**	**cn**	**ar-sy**	**es**	**fa**	**tr**	**en**	**en-zw**
62	34	17	3	3	2	1	1	1
$\sum$	**es-ar**	**ml**	**hi**	**hr**	**it**	**id**	**pt**	**ru**
8	1	1	1	1	1	1	1	1
70								

Table 1: Distribution of mother tongues (L1). German (*de*), Chinese (*cn*), Arabic (Syria) (*ar-sy*), Spanish (*es*), Persian (Farsi) (*fa*), Turkish (*tr*), English (*en*), English (Zimbabwe) (*en-zw*), Spanish (Argentina) (*es-ar*), Malayalam (*ml*), Hindi (*hi*), Croatian (*hr*), Italian (*it*), Indonesian (*id*), Portuguese (*pt*), Russian (*ru*)

3.1. Recording Conditions

The SpeechRecorder software [7] offered by the Clarin-D website was used to collect the PATSY-I corpus. 59 speakers were recorded in a non-noisy environment using a Plantronics headset microphone. The remaining 11 recordings were done mostly in China at the participant's home without additional guidance. To some extent these recordings are affected by background noise and/or bad recording equipment.

Before the recording session, speakers were asked to rate statements about their personality taken from the Big Five Inventory-10 (BFI-10) by Rammsted and John [8] which is a short form of the Big Five Inventory-44 (BFI-44) by John and Srivastava [9]. The BFI-44 describes human personality in terms of 5 factors: "Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience". It consists of 44 statements which are rated on a 5-point Likert scale anchored at 1 = disagree strongly to 5 = agree strongly. To save time, Rammsted and John reduced the number of statements from 44 items in total to 2 items per factor, such that our participants had to rate 10 personality statements. All participants who were recorded for the Patsy-I corpus rated the English version of the statements (both BFI-10 and BFI-44 have also been published in German).

Furthermore, speakers were presented with a form to collect demographic data. The information they were asked to provide was: name, age, gender, place of birth and residence, native language, other languages spoken, English learning method, age of English onset, year of last English lesson, duration of residence in an English speaking country and own English proficiency judgement. Also date, location and recording hardware were collected for each session.

3.2. "Flight Basics" Recordings

To cope with the demands of a phraseology training system for future pilots and ATCOs, we focused on two basic topics of their flight education: call signs and radio (channel) frequencies. Call signs are used for the identification of an aircraft and represent a unique name which consists of a combination of the airline's identification number and the flight number. The ICAO spelling alphabet assigns a unique word to each letter of the alphabet in order to reduce misunderstandings due to poor audio conditions, radio interferences or differing pronunciations. For example, "*a*" is spoken as "*alpha*" and "*z*" is spoken as "*zulu*". For the same reasons, there is also an ICAO pronunciation given for numbers. The "*th*" sound is avoided, such that "*three*" and "*thousand*" become "*tree*" and "*tausand*", "*four*" and "*nine*" are pronounced as "*fower*" and "*niner*". Radio frequencies are six digits long and the period after the first three digits is pronounced as "*decimal*"; for example, 129.775 is spelled as "one two niner decimal seven seven five".

The recording session was structured as follows: first, the participant was asked to read nine sequences of three ICAO spelling alphabet words (e.g. "*bravo romeo mike*"). Then all speakers were shown nine sequences of digits which were between two to six digits long and written as words. After presenting the ICAO pronunciation rules for numbers "*three*", "*four*", "*nine*" and "*thousand*", eleven prompts showing written sequences of the same length as before were displayed. Numbers which should be pronounced in a special way were shown according to the ICAO pronunciation rules (e.g. "tree" instead of "three"). Finally, the speaker was asked to read ten call signs and five frequencies aloud. Digits were written as words and the period was written as "*decimal*".

For each recording, the study participants were shown a prompt which they read aloud. After a fixed number of seconds, the next prompt was shown such that recordings of all speakers were of the same length. Every speaker recorded 44 basic ATC utterances the distribution of which is shown in Table 2. For this part of the corpus, every speaker contributed 4.02 minutes of recorded speech.

	Vocabulary	**Recordings**
Flight Basics	ICAO Spelling Alphabet	9
	English Numbers	9
	ICAO Numbers	11
	Call Signs	10
	Radio Frequencies	5
Number of recordings per speaker in total:		44

Table 2: Number of recordings of type "Flight Basics" per speech task, per speaker.

3.3. "Read Paragraphs" Recordings

In PATSY, the user is asked to answer specific ATC questions in English. PATSY then checks if the input is correct (the right words were uttered) and computes a pronunciation score. If that score is below a certain threshold, the system assumes that the intelligibility of the utterance is not sufficient and asks the user to record it again. To support an improvement of the pronunciation score, the user's utterance as well as the words' correct pronunciation spoken by a reference speaker are played back so he/she can perceive the difference.

When designing the PATSY system, the authors were confronted with the question of whether both the pronunciation and the accent of the user will become more similar to the reference speaker. Levitan and Hirschberg observed this effect which they call "entrainment" for conversational speech. They showed that conversational partners become more similar, for example, in terms of turn-taking behavior [10] or speech phenomena triggering backchannels [11] which listeners use in order to signal continued interest and understanding [12]. For PATSY, users do not take part in a conversation, however, we collected recordings in order to analyse the potential effect of "accent entrainment".

The recording session was structured as follows: first, all 70 speakers were asked to read a part of "The Rainbow Passage" (RR) followed by the paragraph "Please Call Stella" (RS). Then, the participants were shown the prompts for doing the "Flight Basics" recordings described above (see Table 2). Next, the speakers listened to a recording of a reference speaker reading the same part of "The Rainbow Passage" he/she had read before and then they were asked to read it again (L-RR). Finally, for every sentence of "Please Call Stella" the speaker listened to a recording of a reference speaker after which he/she was asked to read it again (L-RS).

Read Paragraphs	**Reading without listening:**		
	"The Rainbow Passage"	(RR)	1
	"Please Call Stella"	(RS)	1
	Reading after listening:		
	"The Rainbow Passage"	(L-RR)	1
	"Please Call Stella"	(L-RS)	8
Number of recordings per speaker in total:			11

Table 3: Number of recordings of type "Read Paragraph" per speech task, per speaker. The ID in brackets behind the name of the paragraph indicates which paragraph was read: "R" for the part of the "The Rainbow Passage" and "S" for "Please Call Stella". The prefix states whether the user read the paragraph without listening to a reference speaker ("R") or he/she listened to a recording before reading the paragraph ("L-R"). In the "L-R" case for "Please Call Stella" each sentence was recorded separately.

The participants listened either to reference speakers with American English (AE) or British English (BE) accent. The AE and BE versions of "Please Call Stella" were taken from the Speech Accent Archive [13]. The AE reference recording and the BE reference recording for the part of "The Rainbow Passage" were taken from the IDEA corpus [14]. All participants of the PATSY recording sessions were either part of the AE or the BE group. As a consequence they were either confronted with reference speakers who had an American (36 participants) or a British accent (34 participants). Eleven recordings were done by every speaker. Table 3 shows an overview with respect to the number of recordings per speaker and reading task. The mean, minimum and maximum recording duration in seconds for every "Read Paragraphs" recording is shown in Table 4. It is interesting to see that on average speakers needed less time when they read "The Rainbow Passage" the second time. The short time span needed for recordings of type "L-RS" is due to the fact that all 8 sentences of "Please Call Stella" were read and recorded separately.

	Duration		
ID	mean	min	max
RR	43.20	31.70	64.15
L-RR	32.86	22.48	67.67
RS	41.28	4.23	88.03
L-RS	6.26	3.40	30.22

Table 4: Statistics regarding the "Read Paragraph" recordings: mean, minimum and maximum duration in seconds.

4. Experiments

In this section we present preliminary results of our experiments on the "ICAO Spelling Alphabet" recordings of the PATSY-I corpus. The goal was to investigate the suitability of the GoP score for automatic pronunciation scoring.

The data used for the experiments comprises 630 recordings (9 per speaker). The computation of the GoP score according to Witt [6] was done using a Julius speech recogniser [15]. The recogniser was trained on recordings of American English speakers and was provided by our cooperation partner e.sigma. We used the recogniser for word and phoneme recognition. The vocabulary of the word recogniser includes all 26 words of the ICAO spelling alphabet and additional pronunciation variants for 18 of these words which results in 54 entries in the pronunciation lexicon in total. For the vocabulary of the phoneme recogniser the 44 phonemes of the English language were used.

The GoP score is a well-known approach for identifying mispronounced phonemes. This similarity measure describes the distance between reference phonemes, based on a recognizer trained with native speech, and the actual phonemes of a user utterance [16], which are in our case the PATSY-I recordings. The score is "0" if there is no difference between the reference phoneme "known" to the recogniser and the phoneme produced by the speaker. A GoP score greater than zero denotes a difference between the actual realisation of the phoneme and its reference.

The mean GoP score was computed for each speaker of the PATSY-I corpus. The statistics on the resulting 70 GoP scores are as follows: a minimum GoP score of 0.41 and a maximum GoP score of 3.35 were observed. The median GoP score was 0.76 and the mean of all GoP scores was 0.87. The great difference between the maximum GoP score and the mean GoP score is due to the high GoP scores of some speakers who were recorded with background noise and/or bad recording equipment as already mentioned in section 3.1. However, within the speakers recorded under lab conditions, the highest GoP score is 1.26.

These results show that the recording conditions have an effect on the GoP score and consequently on the speakers' per-

ceived pronunciation. While this outcome is comprehensible – due to the background noise the accuracy of the phoneme recognizer decreases and the GoP score is worse – it also indicates that differences in recording conditions should be accounted for when comparing GoP scores. However, since these are only preliminary results, this outcome has to be analysed in more detail.

To examine the usability of the GoP score for pronunciation proficiency, we related the computed GoP scores to the self-assessment of English proficiency given by each speaker (see section 3.1). Each speaker therefore rated his/her own English proficiency on a scale of 1 to 6, where "1" represented "Very Good" while "6" was "Very Bad". Values between two integers were also valid. The complete distribution is shown in Table 5. Comparing the speaker proficiency grades of the self-assessment with the computed GoP scores, we obtained a moderate Pearson correlation of $r = 0.49$ (p-value: 0.00002) and a weak Spearman correlation of $\rho = 0.32$ (p-value: 0.00778). These preliminary results show that there is a correlation between the speakers' own perception of their English proficiency and the computed GoP score. However, in order to make more reliable statements about the speakers' intelligibility, we plan on using the word error rate (WER) of a speech recognizer. Another approach on our roadmap is to examine the effect of additive noise on speaker intelligibility.

Self-assessment	1	1.5	2	2.5	3	4	5	6
#speakers	8	2	34	6	15	4	1	0

Table 5: Distribution of English proficiency ratings given by all speakers of the PATSY-I corpus.

5. Summary

The PATSY-I corpus contains speech recordings of 70 speakers of 16 different nationalities. Every speaker made 55 recordings 44 of which contain standard ICAO wordings and 11 recordings are either one or more sentences of the paragraph "Please Call Stella" or part of "The Rainbow Passage". The focus of the PATSY-I corpus is on the pronunciation skills of non-native English speakers based on Air Traffic Control flight communication wording. A key aspect of our approach is to look at the difference between pronouncing English "as the speaker is used to it" and pronouncing English as recommended by the ICAO pronunciation rules, which were designed to increase a speaker's intelligibility. Another goal when recording the PATSY-I corpus was to get data regarding the influence of a reference speaker's accent on the participant. To achieve this goal we made one group of participants listen to a reference speaker with an American English accent and the other group had to listen to a reference speaker with a British English accent before reading the text themselves. In summary, we have on average 4 minutes of ATC flight communication wording and 2.8 minutes of read paragraphs per speaker. We presented a database spanning more than 8 hours of data from 70 different speakers. Our experiments regarding the use of the GoP score for automatically assessing the pronunciation proficiency already show promising results which we will further analyse in the near future. We are also discussing internally the possibility to provide the PATSY-I corpus to the BAS Repository.

6. Acknowledgements

This project was funded by the Federal Ministry for Economic Affairs and Energy's ZIM (Central SME Innovation) program. Some of the recordings used in this project are used by special permission of the International Dialects of English Archive, online at http://wwww.dialectsarchive.com.

7. Bibliography

[1] C. Kaufhold, V. Gamidov, A. Kiessling, K. Reinhard, and E. Nöth, "PATSY — It's All About Pronunciation!" in *in Proc. of Interspeech 2015*, Dresden, Germany, 2015, pp. 1068–1069.

[2] Weinberger, Steven H and Kunath, Stephen A., "The speech accent archive: towards a typology of english accents," in *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Brill, 2011, pp. 265–281.

[3] G. Fairbanks, *Experimental phonetics: selected articles.* University of Illinois Press, 1966.

[4] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." in *in Proc. of Interspeech 2011*, Florence, Italy, 2011, pp. 3081–3084.

[5] J. M. Levis, "Guidelines for promoting intelligibility." in *in Proc. of International TESOL Conference*, Seattle, WA, 2007. [Online]. Available: http://jlevis.public.iastate.edu/intelligibility.ppt, visited 2016-09-15.

[6] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning." *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[7] C. Draxler and K. Jänsch, "SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software." in *in Proc. of LREC 2004*, Lisbon, Portugal, 2004.

[8] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German." *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[9] O. P. John and S. Srivastava, "The Big Five trait taxonomy: History, measurement, and theoretical perspectives." *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[10] R. Levitan, S. Benus, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue." in *in Proc. of AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*, Palo Alto, California, 2015.

[11] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels." in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 113–117.

[12] E. A. Schegloff, "Discourse as an interactional achievement: Some uses of "uh huh" and other things that come between sentences." *Analyzing discourse: Text and talk*, vol. 71, p. 93, 1982.

[13] S. Weinberger. (2015) Speech Accent Archive. George Mason University. [Online]. Available: http://accent.gmu.edu, visited 2016-09-15.

[14] P. Meier, "International dialects of English archive." 1997. [Online]. Available: http://www.dialectsarchive.com/, visited 2016-09-15.

[15] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," 2001.

[16] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation," in *International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 21–30.