

PATSY - It's all about pronunciation!

Caroline Kaufhold¹, Vadim Gamidov², Andreas Kiessling²,
Klaus Reinhard², Elmar Nöth¹

¹Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany

²e.sigma Technology GmbH, Ilmenau, Germany

{caroline.kaufhold, elmar.noeth}@fau.de,

{vgamidov, akiessling, kreinhard}@esigma-technology.com

Abstract

PATSY is an abbreviation for its German name “Piloten/ATC Trainingssystem für den Sprechfunk” which translates to pilot/air traffic control (ATC) training system for radio communication. The phraseology training system is intended to be a stand-alone, platform-independent, multi-user e-Learning framework for learning and practicing the ATC radio communication wordings and at the same time improving intelligibility of the speaker by pronunciation scoring. A serious gaming approach is aimed at, which allows the user to practice his or her communication skills in almost real-life scenarios. At the Show and Tell session at Interspeech 2015, we would like to present a subsystem of PATSY in which we only concentrate on pronunciation scoring. The speaker is prompted to record a sequence of words of the NATO phonetic alphabet and he is given back a visually enhanced feedback regarding his pronunciation score. Further aspects of PATSY are the verification of correct syntax and the assessment of the speaker’s “comfort level” which states how familiar the speaker is with the topic in question.

Index Terms: augmented learning, pronunciation scoring, Computer Assisted Pronunciation Training (CAPT)

1. Introduction

PATSY is our response to the problem of unintelligibility in air traffic control (ATC) communications. Since English is the official language of aviation, every pilot and every air traffic controller (ATCO) is learning the same vocabulary and grammar to communicate with each other. During their education, however, not much attention is paid to the problem of unintelligibility and mispronunciation. PATSY overcomes this gap by teaching pronunciation while learning this “language” which follows a strict and restricted grammar, i.e. <self-id>, <intention> or <self-id>, <confirmation>. The system is intended to accept speech input. As a result, the learner is shown the evaluation of his recording regarding its syntax and a visual representation of its pronunciation scoring.

Topics like the NATO phonetic alphabet which assigns a particular word to each letter of the alphabet, the call sign which is a unique identifier for a flight and which consists of a combination of letters and numbers, or scenarios like arrival and departure, are only some of the topics which are represented as lessons in PATSY. Each lesson contains several units which convey the particular information, make the learner familiar with the topic and test his knowledge as well as his ability to put it into practice given different levels of difficulty. PATSY provides different ways to represent knowledge like text, images, speech and video, and to interact with the learner.

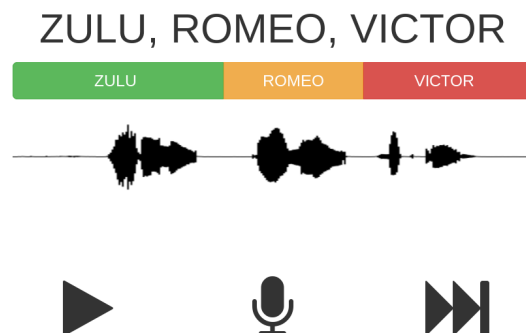


Figure 1: PATSY demonstrator colorizes the recognized words according to their goodness of pronunciation (“ZULU” is good (green), “ROMEO” is middle (yellow) and “VICTOR” bad (red) pronunciation).

The simplest types of interaction are about reading a shown prompt or repeating an auditioned prompt. Thereby the prompt comprises, for example, either some of the letters of the NATO phonetic alphabet or a correct sentence to ask for landing permission. A unit gets more difficult when the learner is asked to substitute the NATO phonetic alphabet for letters and numbers contained in a prompted call sign or if he is asked to complete a radio message asking the ATC for advisory services given only the first part of the sentence. The most difficult type of interactions and the most challenging units for the learner are the serious gaming units which allow the user to fly his own aircraft and demands to master the radio communication without any additional help from the system.

PATSY is intended to not only use speech input for scoring pronunciation, but also to evaluate the degree to which a learner has already fortified the knowledge about the vocabulary and grammar of particular procedures. This shall be done by examining features like the response time or the time when a prompt is displayed until the learner starts to speak. For the Show and Tell session, we want to present our first version of the PATSY subsystem which cares about pronunciation scoring in learning the NATO phonetic alphabet.

The structure of the paper is as follows: Section 2 is giving a short overview of the system architecture of PATSY. The pronunciation scoring subsystem as we want to show it, is ex-

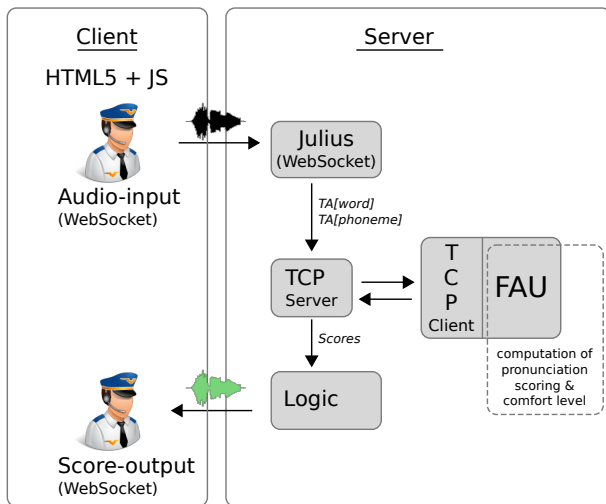


Figure 2: System architecture of PATSY.

plained in Section 3 and planned extensions and topics to improve are listed in Section 4.

2. System Architecture

Figure 2 shows the system components of PATSY: Exercises and learning material is offered to the user by a visually appealing frontend written in HTML5 and JavaScript. Speech input is recorded and then sent via a WebSocket connection to the server. On the server side, a Julius speech recognizer [1] process is waiting for speech input to be processed. The correctness of the speech input is done based on recognition on word level which also provides the time alignment ($TA[word]$). Furthermore, Julius runs an additional phone loop in order to get time alignment information on phoneme level ($TA[phoneme]$). Recognition results are then forwarded to the FAU module over TCP. The FAU module is responsible for computing the pronunciation scores and assessing the user's comfort level. The Logic entity evaluates the scores given by the FAU module and adjusts the server response such that it can be translated by the HTML5 and JavaScript frontend to give visual feedback to the user concerning his assessed goodness of pronunciation and comfort level.

3. The FAU Module

The FAU Module is in charge of computing scores for evaluating the pronunciation and the comfort level of a speaker regarding a particular utterance. As input, the module takes the utterance as WAV file as well as the result of the word and phoneme recognition given by the Julius speech recognizer. Based on this information, features are computed in order to assess the goodness of pronunciation and the speaker's comfort level. As a first approach, the goodness of pronunciation (GOP) score [2] was implemented which computes the ratio between what is expected to be uttered in relation to what has actually been uttered. For near future, we want to extend the FAU Module computations by prosodic [3] and pronunciation features [4]. By incorporating information on rhythm, stress and intonation, we hope to further improve our pronunciation scoring and to gain more knowledge about the prominent features describing differences in pronunciation between speakers of different mother tongue

in ATC communication.

The comfort level is an attribute which we define as the degree to which someone's knowledge about a particular topic has to be fortified in order to finally say that he knows that topic. For example, pilots have to learn how to structure calls to ATC facilities in order to ask for landing permission. An obvious feature for evaluating the comfort level is the time it takes the speaker in order to produce the correct wording to ask for landing permission. For the near future, the incorporation of the assessment of this feature as well as other features into the FAU Module is planned in order to describe a speaker's comfort level. Whereas it will be particularly interesting whether, for example, a low comfort level can be distinctively differentiated from hesitation due to low language proficiency.

4. The PATSY Demonstrator

The PATSY demonstrator which we want to present at the Show and Tell session of Interspeech 2015 is only a subsystem of the PATSY system. It concentrates only on pronunciation scoring and resembles only part of the functionality the FAU Module is intended to provide. As shown in Figure 1, the user is asked to speak a sequence of words of the NATO phonetic alphabet which is in the following referred to as prompt. After recording the prompt by clicking on the microphone icon, the evaluated pronunciation score for every word is shown below the prompt. A button is drawn for each word, whereby its color depends on the goodness of pronunciation given by the computed GOP score for that word. Available colors are green, yellow and red and the evaluation of the goodness of pronunciation is accordingly: good, middle and bad. Until Interspeech 2015, it is planned to also use prosodic features for assessing a pronunciation score and to allow the user to listen to a reference speaker in order to improve his pronunciation in the next recording.

5. Acknowledgement

This research is funded by the Federal Ministry for Economic Affairs and Energy's ZIM (Central SME Innovation) program.

6. References

- [1] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," 2001.
- [2] 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.
- [3] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The prosody module," in *Verbmobil: foundations of speech-to-speech translation*. Springer, 2000, pp. 106–121.
- [4] C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann, "Pronunciation feature extraction," in *Pattern Recognition*. Springer, 2005, pp. 141–148.