A VISUAL FEEDBACK TOOL FOR GERMAN VOWEL PRODUCTION

Patrick Carroll, Jürgen Trouvain und Frank Zimmerer

Saarland University
pcarroll, trouvain, zimmerer [at] coli.uni-saarland.de

Abstract: For non-native German speakers, correct vowel perception can be difficult due to a relatively large inventory of German vowels which are similar in their acoustic qualities and/or duration. For L2 learners whose native language does not distinguish between these similar vowels, several distinct vowels may be assimilated to a single vowel category from the speaker's native language [3]. This perceptual deficit carries over into production, where it has been shown that some L2 German speakers fail to correctly produce German vowel minimal pairs [6]. In order to help non-native German speakers improve both perception and production of acoustically similar vowels, we propose a prototype visual feedback tool which illustrates the differences between the sounds. The tool is based on a listening and repetition exercise, with an added visual modality which provides graphic representation of the first two formants as well as duration. The user can listen to native German productions of target vowels and see formant values and duration visually represented, then record their own productions and see how formants and durations compare. The goal is that with repeated use, the learner can adjust their production of the vowels to approximate acoustic and duration based targets from the native German productions.

1 Introduction

The contents of this paper describe the development a Computer Assisted Pronunciation Training (CAPT) tool intended to improve German vowel pronunciation for non-native German speakers. It will begin with a background of second language learning research which theorizes why certain speech sounds are persistently misclassified by second language learners, and goes on to identify some pedagogic goals for learning new phonetic categories in order to overcome these errors. It then looks at existing CAPT systems to provide examples of successful and unsuccessful feedback strategies which have informed the development of this CAPT tool. The second section describes the system itself, including a description of the user interface, the reasoning behind the feedback method, and the process of vowel detection and measurement. Because the tool is presently still a prototype under development, the third section is devoted to discussing future possibilities. This includes an example of a pronunciation training exercise using the tool, as well as some proposals for improvements, and testing of usability and effectiveness. The final section is reserved for conclusions on the proof of concept of a vowel training system for computer assisted pronunciation training.

1.1 Second Language Learning

Research has shown that adult learners of a foreign language experience persistent difficulty in the ability to perceive certain phonological contrasts due to interference from the phonology of their native language. For vowel contrasts, studies have been performed across several language

pairs which point to difficulties in perception when the learners' native language has a single vowel category in a given acoustic region, and the target language has two or more vowel categories in a similar region. For example, native speakers of Spanish have difficulty perceiving the Catalan /e-ɛ/ distinction [10], and native speakers of Italian have difficulty perceiving the English /a-ʌ/ or /a-æ/ [4]. Difficulty also arises for vowel categories which are contrastive in duration *and* spectral values, such as Italian speakers attempting to discern between the /i-ɪ/ distinction in English [14]. Flege [3] has proposed a theoretical model to account for these phenomena, which hypothesizes that the formation of a new category for an L2 sound may be blocked and falsely classified as equivalent to a category in the L1 if the sounds are similar and learners lack the ability to perceive the acoustic or phonetic features which distinguish between the two. The model goes on to state that the production of the L2 sound will resemble that of the L1 category to which it was assimilated. Work by Hirschfeld [6] in the field of German phonology shows this to be the case, and that foreign learners of German show production problems with several German vowels which do not exist in the their native phonology.

The model by Flege also provides hypotheses as to how new vowel categories can be successfully created. If bilingual learners are able to discern at least some of the features (such as spectral values or vowel duration) that differentiate a sound in the L2 with a vowel category in the L1, they may successfully form a new category. However, the establishment of a new vowel category for bilingual learners may rely on different features or feature weights to define the category boundaries as compared to a native speaker. Experimental evidence of this has been show in the perception of Dutch vowels by Spanish and German learners of Dutch [2]. For example, Spanish learners of Dutch rely heavily on vowel duration to categorize tokens of /a-ɑ/, while L1 German learners and L1 Dutch natives rely more heavily on spectral information. The implication of the research presented in the section suggests that while the phenomena of vowel category misclassification is widespread among L2 learners, errors are specific to the learners L1 background, and strategies for overcoming these errors may also need to be language specific. Using these findings, we identified the following list of priorities for development of a pronunciation tool for vowel segment contrasts: 1.) Learners must be made aware of the features which differentiate similar sounds in the target language. 2.) Information for several features should be provided to accommodate learners of different L1 language backgrounds. 3.) Certain features can be highlighted or emphasized based on the specific language background of the learner.

1.2 Computer Assisted Pronunciation Training

Moving forward from the pedagogic priorities established in the previous section, the focus now shifts to examine how technology can be employed to meet those needs. The development of CAPT is a promising field for language education, and several CAPT systems claim the ability to automatically identify pronunciation errors and provide some form of feedback [9, 13]. This can be a powerful tool in addressing issues with foreign accent which may be difficult or even impossible to isolate in a typical classroom setting [13]. However, there is a great deal of discussion about what parts of the signal are amenable to searching for pronunciation errors (e.g. segment, word, phrase) and once that information is gathered, what are the best methods of providing feedback [5, 9]. We will look at several systems and discuss both advantages and drawbacks of the feedback they provide in hopes of establishing some guidelines for pedagogically sound feedback.

The first type of CAPT system to discuss will be applications which perform some type of automatic speech recognition (ASR) on a user's utterance and return feedback based on the results of the ASR algorithm. A fairly ubiquitous version of ASR based pronunciation training is used

in commercial products which give users a general rating or goodness measure of their accent [13]. These systems have been regarded as problematic by the academic community for two main reasons: First, a single rating on the quality of pronunciation throughout a whole sentence makes little pedagogic sense. The feedback is far too coarse grained to be acted upon, because the learner cannot use that information to improve the production of any specific sounds or the overall prosody of the phrase [5]. Second, the speech recognition algorithm used for evaluating the accent is completely opaque, and therefore it is unclear what qualities constitute a normal accent, and what qualities are considered foreign. One alternative to the general score feedback paradigm is for a CAPT system to employ ASR to identify word level errors. Neri et al. [9] mention two systems which label mispronounced words of a target sentence in red to point out where the user has made a mistake. By making the feedback focused on the word level, the user has a much better chance of being able to identify and correct the error. However, the feedback still remains too implicit, and this leaves the user with the lengthy task of trial and error to arrive at the correct word pronunciation without guidance on what segments or suprasegmentals need attention [5]. While ASR has made great improvements in the past decade in the ability to detect foreign accented speech, at present it still lacks the precision to identify segment level errors, which makes it of limited use for fine grained pronunciation feedback for vowels and consonants [8].

Because automatic error detection remains, at present, computationally difficult, perhaps that task is better suited for the cognitive resources of the user. The CAPT system can instead provide a visualization of what's happening in the speech signal, allowing user identify their own errors and act to correct them. There exists many CAPT systems which do produce some visual analogy of the speech signal (or parts thereof) with varying levels of success [9]. Some early systems provided the user with an oscillogram and spectrogram of their speech production which could be compared to some gold standard. The clear drawback of this type of feedback is that it is very difficult to interpret, and requires prior phonetics training to read the oscillogram and spectrogram [5]. More success has been found in displaying abstract visual feedback derived from the speech signal. For instance a system which display representations of intonation, stress, and rhythm as visual cues to improve pronunciation of suprasegmentals in an L2 language [7], or a system which is designed to correct segment level errors by providing a visual representation of the vocal tract moving with the speech signal [15]. CAPT tools such as the two just mentioned are examples of how automatic signal processing and visualization techniques can provide useful analogies of a speakers voice which would be impossible in a typical classroom setting.

2 System

Examples of CAPT systems previously mentioned have shown that providing useful feedback is challenging, and there are some pitfalls to avoid such as overly generic feedback, or feedback which is difficult to interpret. In addition to the technological considerations, analysis and feedback must be pedagogically motivated with the learner's progress in mind. With these considerations taken into account, the design of this CAPT tool was intended to address the narrowly defined task of vowel perception and production, and to provide feedback to users in a visually intuitive format. Because vowels are difficult to describe in terms of clearly defined articulatory positions, we have opted to adopt an abstract visual approach to feedback which shows the vowel's acoustic characteristics in two-dimensional space, and the duration as a bar graph. By drawing these features, the user has a visual analogy of what happening in the vowel, and can use that as a supplementary source of information when training vowel production.

2.1 System Layout

In what follows we describe the layout and user interface (UI) of the prototype pronunciation training tool. The tool was created using the Praat Scripting language [1] and takes advantage of the functionality of the Praat demo window in order to display the controls and visual feedback regions pictured in figure 1. The region in the bottom left side of the figure (label 1) contains buttons which play back examples of minimal pair nonsense words spoken by a native German speaker. In the example shown, the minimal pairs contrast the categories for the German vowels /i/ and /ɪ/. When a sound has been played, two modes of visual feedback are provided to show the listener some of the the qualities of the vowel produced. Spectral information is provided by plotting a point representing the F1-and F2-values of the vowel (label 6) superimposed on set of targets in acoustic space (label 4). The x-axis of the acoustic space represents F2-values analogous to front and back vowels, and the y-axis represents F1-values analogous to high and low vowels. Target centers are based on average male German productions of the /i/ and /ɪ/ vowels respectively as measured from the Kiel corpus of read speech [11]. The duration of the vowel is displayed to the right (label 5) in green. The user can then attempt their own production of the nonsense word by pressing the record button (label 2). The record function provides by default 3 seconds for the users to record their own voice, after which, the audio is processed and the vowel is plotted in the acoustic space (label 7) and as a duration bar (label 5). Both the plotted point, and the duration bar appear in red and are labeled "user" so that the user can see how their production of the vowel compares to the native German pronunciations labeled in green. The script allows for multiple recordings to be cumulatively plotted in the acoustic space, so that the user can see if new attempts at producing the vowels are getting closer to the targets. The user may also listen to the previous 5 recordings they have produced (label 3) in order to sensitize their ear to the subtle changes in vowel quality which produced different results.

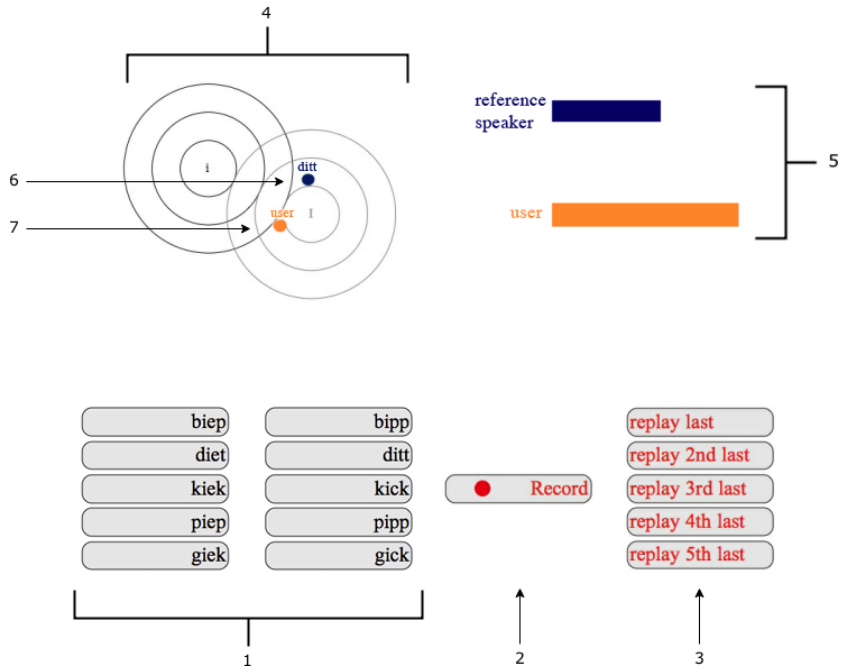

Figure 1 - User Interface with the numbered regions showing the following: 1: playback of native speaker, 2: record button, 3: playback of user, 4; acoustic space with vowel targets, 5: duration space with vowel lengths from native speaker and user, 6 & 7: vowels from native speaker and user plotted in acoustic space.

The design of the user interface is intended to be simplistic and free from extraneous information which might distract the user. For this reason, we have chosen to leave the axis of the acoustic space un-labeled, as well as the duration bars. This decision was made with concern that users seeking a specific number target would be distracted from the larger goal of producing and perceiving native like German vowels. The visual fields of information are thus intended only as a reference to show certain acoustic properties, with the intention that they will contribute to better aural discrimination between two or more similar vowel categories.

2.2 System Design

The technical aspects of system design rely on the functionality of Praat [1] in order to process the signal and provide information for visual feedback. The first task which must be performed is to identify the vowel region in a recording, and create a reference file marking (TextGrid) the beginning and end of the region. This is done by looking at the intensity curve of the recording and selecting the part of the signal which is within 10 dB of peak intensity and which has a duration of at least 40 ms (see figure 2). This approach is a rather naive means of identifying the vowel[1], and relies on the assumption that the vowel will be the most sonorant segment of the utterance and have a relatively long duration. Therefore, in order to control for better vowel identification, all nonsense words used in the training are monosyllabic monophthongs which have short duration, low intensity consonants preceding and following the vowel. After the vowel boundaries have been established, duration is easily calculated, as well as a midpoint for the vowel. From the midpoint automatic measurements of the 1st and 2nd formants are made using a Praat formant object.

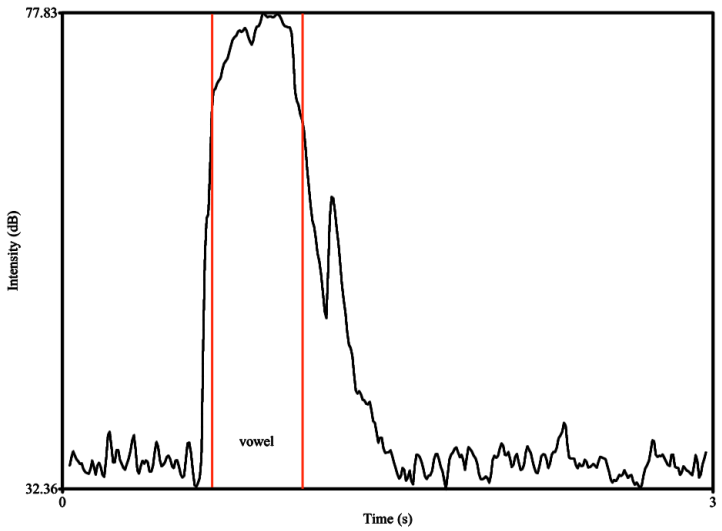

Figure 2 - Selecting the region with peak intensity (Max dB to Max dB -10) as the vowel region.

With the duration and formant information, the script calls a function to plot a point in the acoustic space and draw a duration bar. The same algorithm is used for both the user input, as well as pre-recorded native German input for the sake of consistency. Previous user recordings are also stored for playback, and the buttons in the bottom right side of the UI are updated to reflect the most recent recordings.

[1]For the purposes of prototype development, this method of vowel identification has been effective, however future improvements will be proposed in the following section.

3 Future Work

The vowel training tool remains a prototype, and has not yet been used in an education setting or formally tested. In this section we wish to discuss some thoughts on future work, including an example pronunciation exercise designed with the system in mind. In addition we will cover plans for testing the system's ease of use and effectiveness in improving L2 vowel perception and production. Finally some improvements will be suggested which concerning the systems functionality, and it's general design.

3.1 Example Exercise

What follows is the process envisioned for using the the vowel training tool as part of a pronunciation exercise for a non-native German speaker. Before the exercise itself begins, some initial set up would be required to tailor the task to the user's language background. The user would be asked what their native language is, and asked to read a small set of words with vowel minimal pairs. This information would be used to identify difficulty between German vowel categories and their native phonology, and to select which vowels to train on[2]. After the set-up, the user would be given very simple instructions informing him or her that the area to the right depicts the duration of the vowel, and the area to the left depicts a two- dimensional space showing some acoustic qualities of the vowel. They would then be told that the center of each "bulls-eye" is an approximate target for the acoustic quality of the German vowel depicted therein. Finally they would be told that by changing the shape of their mouth and tongue they can change where a point is plotted in the acoustic quality space, and by changing the length of the vowel, they can lengthen or shorten the bar in the duration space. After this short set of instructions, the user would be given the task of listening to some examples of minimal pairs produced by a native speaker, and then attempting to produce their own. They can use the visual feedback from the acoustic space, and the duration space to experiment with different vocal tract geometries and lengths and see their progress. After the user feels confident they can distinguish between the minimal pairs, they will be given a final post exercise pronunciation test where they will produce minimal pairs without the aid of the visual feedback. This exercise would be repeated on a regular basis and their progress tracked to monitor for improvement.

3.2 Testing

Testing of the tool will need to be considered from two perspectives: effectiveness and usability. To measure effectiveness, the tool must be tested for its ability to improve L2 learners pronunciation of German vowels. L2 learners would be regularly given vowel training exercises (as previously described) over an extended period of time. Recordings of their vowel productions would be saved, and their average duration and formant values would be tracked throughout the training period to see if they are approaching more typical German productions. This would indicate that the L2 learners are successfully creating vowel categories to distinguish between similar German vowels. This group would be compared to two control groups of L2 learners with a similar level of German language learning background. The first control group would receive normal classroom pronunciation training, and the second group would receive no training, so that the tool can be compared to those two baseline outcomes. The second test of the tool would be an evaluation by the L2 learners themselves on how they like the interface. This would comprise of a survey covering each part of the tool, asking how they rate ease of use

[2]At present the system is hard-coded to display the /i/ - /ɪ/ distinction, but future implementations of the tool would ideally have a database of all German vowels produced by several native speakers, and could be set-up to work with any arbitrary vowel distinction.

and intuitiveness. An open comment section would also be provided for suggestions or changes they wish to be made. Using these two means of testing, we would hope to establish if the tool provides a benefit to L2 learners, and if they find it easy to use and understand.

3.3 System Improvements

Through the development process, many areas for improvement were identified which could expand the accuracy and usability of the vowel training tool. The most pressing need is a more sophisticated system for identification of the vowel. The present algorithm limits the tools functionality to measuring only single-syllable words without the context of naturally read or spontaneous speech. For future development of the tool, we propose a forced alignment method of vowel identification which would be able to locate vowel boundaries over the duration of a phrase, rather than a single syllable. This improvement would benefit the user by training them to perceive and produce contrasts in normal speaking contexts rather than under tightly controlled conditions. In addition to more flexible vowel detection, accuracy can also be improved by tailoring the acoustic targets for vowel categories to the individual acoustic space of each user. A relatively easy first step would be to have different targets based on the user gender, as gender is known to affect average formant values [11]. A more ambitious improvement would be to record reference vowels from a user at the extremes of the acoustic space, such as /i/, /a/, and /u/ and adjust the targets of other vowels within the acoustic space relative to those reference points. This would ensure that the users are not trying to imitate an average vowel, but one appropriate to their vocal tract physiology. These suggested improvements would require a dramatic re-design of the underlying code for the tool, which would ultimately be beyond the scope of Praat scripting. Therefore a future version of the tool would likely be designed in a more flexible language, such as Python [12].

4 Conclusion

The current goal of development for this vowel pronunciation training tool has been to demonstrate a proof of concept for providing a useful and intuitive feedback method in a CAPT system targeting segmental errors. Based on the pedagogic theories outlined in section 1, we have concluded that L2 learners must learn to distinguish which features separate a new foreign vowel category from a similar category in their L1. Because audio playback of similar vowel may not be an effective teaching method on it's own, we assume that including a visual representation of certain features may aid L2 learners in understanding the difference between similar vowel sounds, and ultimately help them produce better examples of these vowels themselves. Visual feedback was also intended to be as straightforward as possible by showing changes in features as essentially distances and quantities. This was intended to allow the learner to quickly begin experimenting with the tool, and to free them from the burden of having to learn specific background knowledge of phonetics or phonology. We believe that this focus on simple feedback, and learner guided experimentation will allow them to identify their mistakes, and act to correct them. While there remains a great deal of work to be done in improving the design and testing its effectiveness, the initial challenge, of measuring a vowel and displaying it for feedback, has been met.

References

[1] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer*. Computer program (Version 5.3.64), 2014.

[2] ESCUDERO, P., T. BENDERS and S. C. LIPSKI: *Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners*. Journal of Phonetics, 37(4):452–465, 2009.

[3] FLEGE, J. E.: *Second language speech learning: Theory, findings, and problems*. Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research, pp. 233–272, 1995. Timonium, MD: York Press Inc.

[4] FLEGE, J. E., I. R. MACKAY and D. MEADOR: *Native Italian speakers' perception and production of English vowels*. The Journal of the Acoustical Society of America, 106(5):2973–2987, 1999.

[5] HANSEN, T. K.: *Computer assisted pronunciation training: The four'K's of feedback*. Current Developments in Technology-Assisted Education, pp. 342–346, 2006.

[6] HIRSCHFELD, U.: *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*. Frankfurt am Main: Hector, 1994.

[7] KOMMISSARCHIK, J. and E. KOMISSARCHIK: *Better Accent Tutor–Analysis and visualization of speech prosody*. Proceedings of Internation Speech Technology In Language Learning 2000, pp. 86–89, 2000.

[8] NERI, A., C. CUCCHIARINI and H. STRIK: *Automatic speech recognition for second language learning: How and why it actually works*. In *Proc. Internation Congress of Phonetic Sciences*, pp. 1157–1160, 2003.

[9] NERI, A., C. CUCCHIARINI, H. STRIK and L. BOVES: *The pedagogy-technology interface in computer assisted pronunciation training*. Computer Assisted Language Learning, 15(5):441–467, 2002.

[10] PALLIER, C., L. BOSCH and N. SEBASTIÁN-GALLÉS: *A limit on behavioral plasticity in speech perception*. Cognition, 64(3):B9–B17, 1997.

[11] PÄTZOLD, M. and A. P. SIMPSON: *Acoustic analysis of German vowels in the Kiel Corpus of Read Speech*. Work reports of the Department of Phonetics and Digital Language Processing of the University Kiel], pp. 215–247, 1997.

[12] PYTHON SOFTWARE FOUNDATION: *Python Language Reference*. Computer program (Version 2.7), 2014.

[13] THOMSON, R. I.: *Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation*. Calico Journal, 28(3):744–765, 2011.

[14] WEBER, A., A. M. DI BETTA and J. M. MCQUEEN: *Treack or trit: Adaptation to genuine and arbitrary foreign accents by monolingual and bilingual listeners*. Journal of Phonetics, 46:34–51, 2014.

[15] WONG, K.-H., W.-K. LO and H. MENG: *Allophonic variations in visual speech synthesis for corrective feedback in CAPT*. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5708–5711. IEEE, 2011.