AzARc – A Community-based Pronunciation Trainer

R. Kompe, M. Fischer, D. Hirschfeld, U. Koloska, I. Kraljevski, F. Kurnot, M. Rudolph

voice INTER connect GmbH
Ammonstraße 35, D-01067 Dresden, Germany
{kompe,fischer,hirschfeld,koloska,kraljevski,kurnot,rudolph}@voiceinterconnect.de

Abstract

In this paper a system for automatic pronunciation training is presented. The users performs exercises where their utterances are recorded and pronunciation quality is evaluated. The quality scores are estimated on phonemic, prosodic, and phonation level by comparison to the voice quality parameters of a reference speaker and presented back to the users. They can also directly listen to the voice of the reference speaker and try to repeat it several times until their own pronunciation quality improves. The system is designed in a flexible way such that different user groups (disabled, children, healthy adults, …) can be supported by specific type of exercises and variable degree of complexity on exercise and on GUI output side. One focus is on a client-server based realisation, where the database stores exercises as well as the results. Therefore therapists, (language teachers) can monitor the progress of there patients (pupils), while they train by themselves at home, and, moreover, anybody can introduce his own set of exercises into the database and is free to make them available to other people on a win-win basis.

Keywords: computer assisted pronunciation training, automatic speech quality measurement, open database with speech training exercises

1. Introduction

In the past we have developed a system, called AzAR, which is a PC-based personal tool supporting foreign language learners to improve their pronunciation (akzentfrei-sprechen.de, currently only German version is available). With this system the users performs exercises presented on the screen, speaking words, word-pairs, phrases or longer texts.

The user's pronunciation and the prosody (i.e., pitch, loudness contour and relative duration) are evaluated using standard speech recognition technology. The speech signal is segmented automatically on phoneme-level and the pronunciation quality is marked by colors for each phoneme. In addition to the speech signal and the phoneme segmentation, also the spectrogram, the pitch and energy contours, and the formant trajectories can be visualized. Play-back with different speed for arbitrary portions of the user's utterance or of the one of the reference speaker is possible as well. Furthermore, an animation of the lip rounding and movement as well as the whole vocal tract is integrated in the feedback interface. This can be watched like a video for the whole utterance or for individual phonemes.

This system about to be re-implemented and expanded, so that it can be used as a pronunciation trainer for a broad range of user groups, e.g. for speech impaired people (after stroke or Parkinson disease), for children or just for those who often hold presentations or give speeches and who want to improve their pronunciation.

This requires an improvement and extension of the exercises, which has to be reflected in an extension of the algorithmic base for pronunciation quality measurement and the capabilities of the user interface. In the case of people having a speech impaired conditions (disease), a therapist would supervise the exercises off-line.

The system architecture is based on a client-server-based solution. The user interface runs on a mobile client like Android tablet, iPad or even smartphone as well as on PC. The interface, database and the application server are developed by our project partner tyclipso (tyclipso.net). A theoretically arbitrary number of simultaneous client connections at the same time to the server is possible. The server receives the recorded speech signals, executes the audio processing and compares the results with the ones computed from the reference speech data.

The results and the scoring are sent back to the client and presented in an adequate way, taking into account the user's cognitive capabilities and his assumed knowledge about speech, language and phonetics. The server also keeps an anonymous client database to record the exercises done over time including the scoring that indicates the learning progress.

The server automatically provides the user with the next suited exercise according to the learning progress and previous history using some heuristics or based on a set selected by a therapist for specific patient. Therapists always have access to all results and speech data of their patients for examination and can anytime adjust the set of exercises.

This paper describes in more detail each of the aspects of the new system, and it is organized as follows, the Section 2 gives the description of the exercise database, next section presents the typical workflow of the system. User interface concepts are given in Section 5 and audio processing methods are presented in the following one. Pronunciation quality measures are listed in Section 6, the learning progress in 7 and the overall system architecture is described in Section 8. Section 9 is focused on the community content database and its characteristics. The paper is concluded with the Summary section.

2. Exercise Database

The various user groups have very different needs in the type of exercises. We have put together a large set of exercises reflecting these different needs including breathing exercises as well as training of mimic and motor functions and coordination of mouth muscles. Exercises concerning phonation, prosody, and articulation

start with single phonemes, continue with words and word pairs, ending up in long sentences or short stories depending on the maturity level of the user. In the following, we give only a few examples for persons with speech impaired conditions:

Loudness: difficulties to speak aloud. Therefore, users are asked to speak isolated vowels or words (e.g., "stop"), short phrases like "good morning" very loud. Also sentences can be presented to the user, where the key issue is either to speak with a loud voice constantly from beginning to end of the sentence or to pronounce every n-th word aloud.

Voicing: trouble with voicing of any phoneme. There voice sounds more or less like whispering. This can be trained by a wide variety of stimuli where the degree of voicing is measured on a phoneme-by-phoneme basis.

Pitch: problems with articulation of certain pitch levels at all. Simplest type of exercises is to pronounce different vowels in isolation with low, high or mid-level pitch. Next step is to raise the pitch on a vowel or to produce a falling intonation, or to produce high-low-high pitch patterns or vice versa.

Accentuation: People learning a foreign language often put word and phrase accents on the wrong position. People with certain types of diseases speak rather monotonous. In order to train the accentuation users are presented in the beginning with word groups where they are asked to put the accent on different words when repeating the word group, e.g.:

"the **fast** car – the fast **car**"

More advanced users will be presented sentences containing a few words, and are asked to repeat the sentence several time, each time putting the main accent on a different word. People learning a foreign language in particular need to learn to put the accent on the right word within a phrase embedded in some context. For this, phrases or sentences expressing a contrast are useful, e.g.:

"the black pen – the **red** pen" or

"In our garden, besides of fruit trees and vegetables there grow as well some **beau**tiful **ros**es."

Also, poems are well suited to train people towards a more expressive accentuation.

Articulation: Most exercises focus on training the pronunciation of one specific phoneme. There are some single word exercises, but in most cases, articulation is trained with minimal pairs containing certain articulatory qualities, e.g.:

"**s**eal – **z**eal" (voicing),
"b**ea**n – b**i**n" (vowel length),
"s**a**le – s**oi**l" (phoneme class).

Also we have collected a lot of exercises where only the phoneme to be trained is kept constant and all context is varied, e.g.:

"**s**un – **s**in – **s**ick – **s**ad – **s**oul – **sc**ene – **s**oup – **s**ort".

We have introduced as well exercises with which consonant clusters can be trained; some German examples with pronunciation given in SAMPA alphabet:

"**Kr**opf" /**k r** O pf/
"**Str**asse" /**S t r** a: s @/
„**Pfr**ünde" /**pf r** Y n d @/
„Geri**chtsr**at" /g @ r I **C ts r** a: t/

Phoneme articulation can also be well trained when in a word group every word starts with the same phoneme, e.g.:

"**l**oving **L**atin **l**iterature"

Also, our set of exercises contains a number of sentences, which are designed such that a specific phoneme occurs very often. Tongue twisters are an excellent exercise for advanced users. For further reading concerning training pronunciation, please refer to Celce-Murcia et al. (1996) or Fiukowski (2004).

For each of the most frequent user groups, individual exercises are grouped together for a default training plan.

However, in particular, therapists wish to select a very specific set of exercises meeting the training needs of a particular patient. Therefore, we have designed a large set of exercises each of them annotated with attributes characterizing the exercise. Users will be able to search for specific exercises or to browse through a list of exercises of a certain broad category. Such attributes concerning exercises for articulation, just to give an example, include:

- the position of the phoneme to be trained in the word (initial, mid, final),
- the phonemic context specified by individual phoneme name and phoneme class (e.g., fricative),
- the number of words and syllables in the exercise,
- whether the exercise consists of a minimal pair and which type of minimal pair (e.g., short/long vowel),
- whether it is a phrase or a complete sentence, etc.

As for accentuation, among others, important criteria are whether the vowel to be accentuated is

- short/long or
- open/closed.

The design of the exercises for the first field test was done by psychologist with support of a logopedics therapist from the Bavaria Klinik, Kreischa.

3. Typical Work Flow

Usually at the beginning, training or therapy plan for several days or even weeks is created. A therapy/training consists of several training units, each taking usually 15 to 45 minutes depending on user's abilities. Training units are scheduled such that there is usually only one session per day.

Each training unit consists of a number of exercises ideally to be performed. In most cases users would perform the exercises of a single training unit consecutively, but pauses of arbitrary length are possible. Experienced users may be allowed to select, repeat or skip unfinished exercises by themselves (or for their patients), other users will be offered exercises one after another determined either by some heuristics by the system or according to a list of exercises predetermined by a therapist.

To start a training unit, the user has to logon to the system on client side. Afterwards the client connects to the server and passes the user identity to the server. All communication between client and server takes place using a secure transfer protocol. The server knows about

the user and the status of his training plan, i.e., the exercises performed so far together with a quality scoring for each exercise. Based on this status information, the exercises for this training unit are presented to the user one by one. Depending on the type of user, new exercises are presented or certain exercises from the previous training units are repeated.

When an exercise, e.g. a phoneme or text, is presented on the screen of the client, the user is requested to perform this exercise. His speech signal is transferred to the server while speaking and audio processing and quality measurement starts immediately with the first blocks of audio data coming in (streaming). Results are transferred back to the client as soon as they are available. Certain processing modules deliver results for each 10 ms frame, some with delay, others almost immediately. Other processing modules only deliver results after a certain number of phonemes have been spoken or even just after the end of the utterance.

The results of audio processing and quality scores (meta-data) are sent in "raw" format to the client. The client then decides how to present these results to the user (see Sec. 4). For training pronunciation it is most important that the user can listen to a reference speaker and try to imitate his voice. Therefore, our quality scoring is not based on some average statistical models but on the particular reference speech signal just in use (see Sec. 5). A user may playback the whole utterance or arbitrary segments of his own or the reference utterance.

The same exercise might be repeated several times in a row or on another day. It might be repeated until some quality level or a maximum number of repetitions is reached. Any user is allowed to reject exercises, e.g., because he considers them as being too difficult.

A training unit finishes either after all exercises of this unit have been completed or after a time-limit has been expired. In particular people with speech impairment should keep to a pre-defined time limit whether the number of exercises has been finished or not.

4. User Interface Concepts

The user interface must be flexible in order to adapt to the capabilities of the hardware (e.g., tablet versus PC), being easily portable to various operating systems and screen sizes, and most importantly it must adapt to the maturity level and physical or cognitive capabilities of the user.

An expert like a therapist can be presented with very detailed information (phoneme segmentation, spectrograms, formants etc.) of all the different processing stages. In order not to overload the screen, different views with subsets of available information can be selected. Children, elderly or disabled people need a screen with minimum of information necessary to fulfill the exercise and to provide feedback on the quality or success. Even more, these groups of users must be motivated by some animated feedback, e.g., they could move a balloon up and down by moving their pitch level up and down.

Feedback of the success will also be given in very different form. For example concerning articulation mature users will be presented with the phoneme segmentation and some quality score representation (e.g., a level of colors from red to green) for each phoneme.

Other users will be presented with the text, which was to be read, and the phonemes with worst quality will be mapped to the letters, which are then marked by some color. Users then can get, if they wish, for each such markers a verbal explanation of how they could improve.

Such suggestions look like:

"Please, raise the tone more strongly"

"Articulate this more clearly"

In the future we might be able to give more precise hints like:

"Open your mouth wider" or

"Try to bring your lips in a rounder form"

For users with severe conditions, feedback like this:

"congratulations, you have completed this exercise."

could be given.

5. Audio Data Processing

For each type of exercises different audio data processing is necessary. Some exercises focus on the articulation of phonemes, others on the intonation or the distribution and quality of accents within an utterance, the purpose of others is just to measure the loudness of the speaker's voice and the strength of voicing of to-be-voiced phonemes. In the exercises database, each of the exercises is encoded with information which set of processing modules is required, and how the results should be presented to the user. The exact form of presentation takes into account also the specific needs of the user.

The user's speech has to be processed in real-time manner, hence the audio processing algorithms should provide immediate results where it is possible. The following base processing modules are implemented: normalized loudness (two methods), spectrogram (for direct display and as input to some of the other modules), pitch, accentuation, degree of voicing, jitter, shimmer, formant trajectories and phoneme segmentation. The phoneme segmentation component is based on the Pocketsphinx speech recognizer (Pocketsphinx, 2013) and its Gstreamer version; it is chosen because of the low processing and memory footprint, Many clients will connect at the same time and many instances of Pocketsphinx might be running in parallel.

Because of the real-time constriction, each speech frame is processed as soon as it is available. This reduces the choice of the available algorithms, providing, in some cases, non-optimal estimation of the acoustical parameters. This is the case for the pitch and formant contour estimation, where in the first LPC cepstrum is used for frame-based pitch determination, similarly for the formant trackers. Median filtering is applied for contour smoothing and reduction of the outliers.

Each of the processing modules operates in real-time providing results with the lowest possible latency, except for the phoneme segmentation. Here, the whole utterance should be available to produce final recognition result presented with phoneme boundaries and likelihood scores required for the pronunciation scoring. The acoustic models used for the phonetic alignment were trained on the Verbmobil I spontaneous speech corpus (Burger, 2000). The speech recognizer is configured for phoneme recognition, where the "pronunciation dictionary" is faked to consist only of phonemes. Forced-

alignment is performed by means of using separate finite-state grammar for each of the exercises containing a single state sequence without alternatives which corresponds to the words to be spoken.

As well for the detection of accents the speech signal of the whole utterance needs to be available. Therefore, we can apply a more time consuming multi-channel pitch detection algorithm with more sophisticated smoothing as described in (Noeth, 1991). Position and strength of phrase accents are computed as described in (Kompe, 1997). All the estimated parameters are combined in one data stream and sent back over the network to the client side for synchronous and real time presentation.

The same processing components are used on the utterances of the reference speaker to build a database with acoustical parameters, they are used for comparison and scoring against the user's speech.

6. Quality Measurement

We compare the users speech to the one of a reference speaker rather than to a (statistical) model being optimized on a large set of speakers. By this we want to resemble a training conducted by a teacher or a therapist, where the most important element is that the client/patient listens to the referee voice of his trainer and try to repeat it as close as possible.

For articulation scoring, two options are available, log-likelihood scores obtained from phoneme and forced-alignment recognition, with and without comparing with the referent sentence (Witt et al., 1997). For prosody (global and segmental) scoring, pitch and energy contour matching against the referent sentence using some comparison method (normalized correlation coefficients or Dynamic Time Warping measures). Duration scoring can be estimated for each pair of user and referent sentences, word, phoneme and silence duration are determined from the phonetic labels derived by the forced-alignment recognition. Also, mean values and variances of the F0, formants, intensity, as well as the values of jitter and shimmer are taken into account. These measures can be combined in more general pronunciation quality score and presented back accordingly to the user or the therapist.

7. Learning Progress

For users and therapists it is important to get an idea of the learning progress.

XXX Presentation to the user / therapist: average of gradual scores per unit / change over time – or, “congratulations, you have completed 10 more exercises”.

XXX Some patients, e.g. having Parkinson’s disease, will usually not have a learning progress in the sense of improvement. The goal of a therapy is to slow down the decline in pronunciation abilities. They usually will be treated with a small set of exercises repeated in every training unit. It is important to give also these people some motivating feedback; however, it cannot be in the form of concrete progress.

XXX Computation (research issue?)

8. System Architecture

We decided for a client/server based solution in favor of a personal software solution, because:

- it provides more flexibility with respect to sharing the database of exercises among users,
- new audio processing modules can be more easily added to the system – even on the fly, since we use the Gstreamer plug-in architecture, and
- therapists can monitor and track the progress of patients and adjust the set of exercises for the upcoming training units based on the past training progress without the need for the patient visiting the therapist.
- Mobile devices don´t provide the necessary computational power for upcoming audio data processing, which might be more complex than the ones implemented currently.

All processing modules for the audio stream are implemented as such Gstreamer plug-ins on server side, except for the pitch computation, which in addition runs on the client side, because exercises concerning the intonation, like holding a high/low tone or producing rise-fall intonation patterns, need feedback to the user within 50 ms, which cannot be guaranteed when processing is done on the server.

The overall system architecture is depicted in Fig. 1. The client interacts with the user and transmits the audio signal to the server side and receives the processing results as well as the exercises and reference data for presentation to the user. The media server (MS) receives the audio signal from the client and does all the processing and quality measurements. The application server (AS), holds a user database and a database with exercises and reference data. It is responsible for all session management.

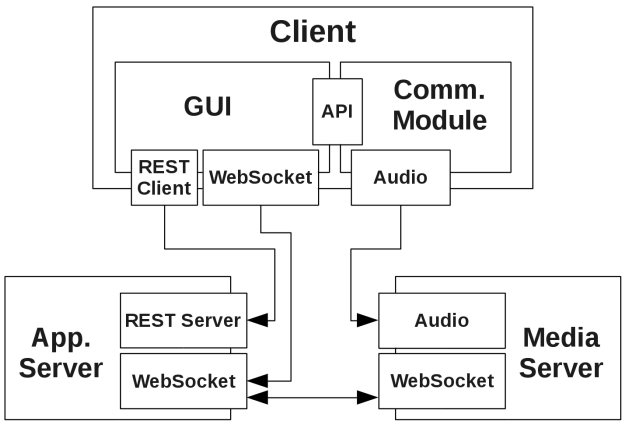

Fig 1: Overall system architecture.

The client consists of four modules: GUI and application handling, audio communication and processing. When a user entered his password, the client establishes a REST and a WebSocket connection to the application server. The REST connection allows the client to call functions on the server, which is necessary, for example, when asking for the next exercise or for login procedure. Furthermore, it provides flexibility concerning the responses to a request. A request opens a channel, which stays open such that a series of response

messages upon one request are possible. The WebSocket channel is used for transmitting messages from AS to client. In particular, the results of the audio signal processing and some graphics pre-computation are transmitted to the client via this WebSocket connection, which is established just once by an http-request at the beginning of a training unit and provides an always open, bi-directional communication connection.

As well, AS and MS communicate through a WebSocket channel. The AS verifies the user ID and transmits the data for user identification to the MS. It then sends the first exercise to the client and at the same time relevant information concerning the required audio processing modules to the MS. The client establishes a secure PJSIP-based communication channel to the MS.

The MS accepts the connection if the user ID can be verified by the information received from the AS beforehand.

The user may now start recording of his speech. The communication and application module retrieves the audio signal from the microphone, feeds it directly into the pitch processing module, if required by the exercise, and, in parallel, sends the audio signal via the PJSIP connection to the MS. The MS internally calls the required audio processing modules as soon as some audio data comes in, does a multi-plexing of the asynchronous processing results and sends one stream of results to the AS.

The AS forwards these results immediately to the client and at the same time stores them together with the exercise ID and the audio signal in the user database. With this user database the AS and eventually the user himself or his therapist can keep track of the training and learning progress. In particular a therapist in theory has the opportunity to check each single exercise the user has performed.

The AS maintains as well the content database, which stores exercise together with reference data. It will be explained in more detail below.

All messages transferred over the Internet are encrypted using secure protocol standards. In the case of the audio transmission, we use SRTP.

9. Community Content Database

The AS maintains a database, which stores all the contents needed to perform an exercise and to judge upon the quality of performance by the user. The latter comprises the speech signal produced by a reference speaker and some additional data like pitch contour, phoneme alignment, accent markers.

Schools teaching pronunciation or therapists might wish to add exercises designed by themselves meeting their specific needs or flavor of teaching/performing a therapy.

Therefore we intend to open the database for anybody to input his own content. As a minimum the task, e.g., the text to be spoken, and the reference utterance must be entered into the database. The user then will be asked to set the attributes specifying the type of the exercise (see Sec. 2). Additional data like pitch or phoneme alignment will be computed once by the media server and stored as well in the content database. People are free to restrict the use of their data just for their own training purposes, i.e., just making it available to their clients/patients. But most beneficial will be if such data is exchanged with other users of the system. This can be done in exchange for other training data or for a fee. Exact policies are still to be defined.

For data to be exchanged we plan to establish a quality measurement service, which certifies the quality of the training material prior to the exchange with other people.

Since our quality measurement method compares the user's utterance directly with the one of a reference speaker, this method is rather effected by dialects, sex, and age of speakers. Therefore, we will offer the possibility to link different reference speech data to each exercise based on such criteria. This in particular might make the exchange of training data via our database valuable.

10. Summary

This paper presents conceptual and architectural overview of a community based computer aided pronunciation training system. It is intended for a broad range of different user groups (children, speech impaired, language learners etc.). The system is based on client-server architecture, the users interact over graphic interface customized according their needs on the client where the server side performs the computational intensive speech processing and scoring in real-time. The scoring is performed against referent speech data in terms of articulatory and prosody quality. The exercises and the referent data are stored in a community content database. The database content can be exchanged by users, language teachers, therapists.

References

Burger, S., Weilhammer, K., Schiel, F. and Tillmann, H.G. (2000): Verbmobil Data Collection and Annotation. In *Verbmobil: Foundations of Speech-to-Speech Translation* (Ed. Wahlster, W.); Springer; Berlin/ Heidelberg.

Celce-Murcia, M., Brinton, D. M. and Goodwin, J.M. (1996) *Teaching Pronunciation. A Reference for Teachers of English to Speakers of Other Languages.* Cambridge: Cambridge University Press.

Fiukowski, H. (2004): *Sprecherzieherisches Elementarbuch*. Niemeyer, Tuebingen.

Kompe, R. (1997). *Prosody in Speech Understanding Systems*. Heidelberg: Springer.

Nöth, E. (1991). *Prosodische Information in der automatischen Spracherkennung - Berechnung und Anwendung*, Niemeyer Tübingen.

Pocketsphinx (2013). Speech recognition toolkit: http://cmusphinx.sourceforge.net/

Witt, Silke, and Steve Young. *Computer-assisted pronunciation teaching based on automatic speech recognition.* Language Teaching and Language Technology Groningen, The Netherlands (1997).

This work is being funded in part by the German Federal Ministry for Economy (BMWi) under grant no. KF2403503BZ2.