# QUALITY ASSESSMENT IN PRONUNCIATION TRAINER FOR SPEECH DISORDER THERAPY

*I. Kraljevski, R. Kompe, R. Jaeckel*, G. Strecha*,*
*F. Kurnot, M. Rudolph, D. Hirschfeld, R. Hoffmann**

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

**TU Dresden, Chair for System Theory and Speech Technology, Dresden, Germany*
*{rainer.jaeckel, guntram.strecha, ruediger.hoffmann}@tu-dresden.de*

**Abstract:** In this paper, a Computer-Aided Speech Therapy system for dysarthric speakers is presented. It is based on a client-server architecture, supported by community-based multi-media database allowing content sharing and exchange. The speakers perform exercises where the pronunciation quality of their utterances 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. The client-server based implementation, allows the therapists to monitor the progress of their patients, while they train by themselves at home. For testing and evaluation of the speech processing and scoring concepts, a domain-specific speech corpus was recorded, over which the dependency and the optimal contribution of prosody and articulation scores was statistically assessed. The statistical test confirmed that in general scoring significantly differ between the healthy and dysarthric groups.

## 1 Introduction

Speech impairment is the inability to produce normal speech due to difficulties in articulation and intonation. Difficulties in pronouncing the sounds or in voicing, in general are some examples of speech impairment. Speech difficulties can also be associated with cerebral palsy, hearing impairment and brain stroke or injury. People with speech impairments have difficulty in expressing their ideas which disrupts their ability to communicate. The most common disorders associated with speech and language production are Aphasia, Dysphonia, Dysarthria or Apraxia of speech. People with nerve or brain disorder have the inability to control the larynx and the vocal cords, which causes the condition known as Dysarthria, resulting in improper pronunciation and speech with low intelligibility. The pronunciation of dysarthric speakers often deviates from that of non-dysarthric speakers in several aspects: lower speech rate, differently pronounced segments, less consistent pronunciation and for longer stretches of speech, pronunciation can be even more varying due to fatigue. Speech therapy with these patients can improve their speech production skills and consequently the intelligibility to some extent, but still oral communication might remain difficult. Computer-aided speech therapy systems are getting more reliable and affordable to speech therapists and people suffering speech impairments. Such systems take advantage of using speech technologies, particularly Automatic Speech Recognition (ASR). ASR was used for recognition of dysarthric speech [1] and in speech therapy applications. Speech quality assessment by an ASR is important for research and clinical purpose in order to determine the functional capabilities of speech and voice production [2]. Several research studies observed that just by using common ASR systems frequently, dysarthric speakers can improve their intelligibility [3]. The ASR component can also be employed in systems specialized for pronunciation training similar like those developed language learning (CALL) and pronunciation training (CAPT) [4][5].

This paper presents a specialized pronunciation training system for (but not exclusively) dysarthric speakers based on the concept of an earlier developed system, called AzAR, a PC-based personal tool supporting foreign language learners to improve their pronunciation [6]. The pronunciation scoring concepts were tested on a recorded corpus of German native speakers divided into reference, test and control groups. The statistical tests showed that the control group achieved better averaged pronunciation scores than the test group, which was expected and support the proposed scoring scheme. The paper is organized as follows: Section 2 gives a description of the proposed framework, Section 3 presents the overall system description and the experimental setup. The quality scoring concepts are described in Section 4 and results and the discussion are presented in the following one.

## 2 Proposed framework

The pronunciation trainer is based on a client-server architecture which gives the possibility of using, besides personal computers, various mobile clients, like tablets or smartphones. A theoretically arbitrary number of simultaneous connections of clients 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 possibility for content sharing and exchange between users and experts is supported by community based multimedia database. The system can be used as a pronunciation trainer for a broader range of user groups, not only speech impaired persons, but anyone who want to improve their pronunciation. This requires an enhanced set of exercises, which has to be reflected in an extension of the algorithmic base for pronunciation quality measurement and the capabilities of the user interface.

The patients perform exercises presented on the device display, uttering words, word-pairs, phrases or longer texts. The pronunciation and the prosody measures are compared with the ones of a reference speaker by means of standard speech technology. The speech signal is automatically segmented and the basic parameters are calculated in real-time manner. The pronunciation quality can be visualized on the subsegmental and suprasegmental level. For example, the pronunciation goodness can be marked by colors for each phoneme along the visualization of the spectrogram, pitch, short time energy and the formant trajectories.

The results and the pronunciation scores 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. Each patient will be provided 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. Therapists always have access to all results and speech data of their patients for examination and can anytime supervise and adjust the set of exercises (on- or off-line).

The system aims to help in developing or correcting the speech (articulation, intonation, loudness, rhythm etc.) of patients with speech impairments. Therefore, it is very important that the system presents the speech parameters in a way that is understandable for the patients while remaining correct from the acoustic-phonetic point of view. Various patients groups have different needs regarding exercise type. A large set of exercises reflecting these different needs was compiled, including breathing exercises, as well as training of mimic and motor functions and coordination of mouth muscles. Exercises concerning phonation, prosody, and articulation starts with word-level (one syllable), continue with word pairs, ending up in long sentences or short stories depending on the progress level of the user [7] [8]. For each of the most frequent user groups, individual exercises are grouped together for a default training plan. The therapists, in general, would like to select a very specific set of exercises meeting the training needs of a particular patient. Therefore, a large set of exercises is designed and each of them is annotated with attributes characterizing the exercise. Users will be able to search for specific exercises or to browse through a list of a certain broad category.

For the patients and the therapists it is important to get an idea of the learning progress. The first measurement is how many exercises have been completed at all and the success percentage. This can be broken down to type of exercises. A further measurement is the extent of improvement of quality scores, which is only meaningful if the same exercises are repeated after some time. Some patients, e.g., having Parkinson's disease, will not have a learning progress in the sense of significant improvement of pronunciation quality. The goal of a therapy is, at least, to slow down the decline in pronunciation abilities. They usually will be treated with a small set of exercises repeated in every training unit.

# 3 System description

## 3.1 Speech corpus

Domain-specific corpus was recorded for the purpose of preliminary investigation of speech production in speech impaired patients, specialized on pathological speech and pronunciation errors. Also, the database was used for testing and evaluation of the speech processing components as well as the quality scoring concepts. The recorded subjects are German native speakers divided into referent, main and control group. The control group consists of 3 healthy speakers, the main group has 25 patients with Parkinson disease and brain stroke, and the referent group has one male and one female speaker (in total 30). The main and control group were recorded in similar technical conditions in office or room environment. Textual data, list of words and the sentences were provided by linguistics experts, as a test stimuli for the recording process, divided into several categories according to the intended exercise.

## 3.2 Speech processing methods

For each type of exercises combination of different speech processing methods is required. Some exercises put more focus on the articulation of phonemes, others more on the intonation or the distribution and quality of accents within an utterance, the purpose of some 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 about the set of required processing modules, and how the results should be scored and visually 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: short time energy, spectrogram, fundamental frequency (F0) track, formant trajectories and phoneme segmentation.

On Figure 1, the processing framework and the separate modules providing the feedback to the users are presented. The incoming data stream, consisting of speech and meta-data (exercise identifier) are separated and the speech is processed in real-time parallel analysis modules. Each module produces a data stream with predefined binary format. The resulting binary streams are analyzed and the articulation and intonation quality scores are estimated and multiplexed with the outgoing streams for efficient network transport. On the user's side, the feedback data streams can be uniquely extracted and accordingly presented to the patient/therapist.

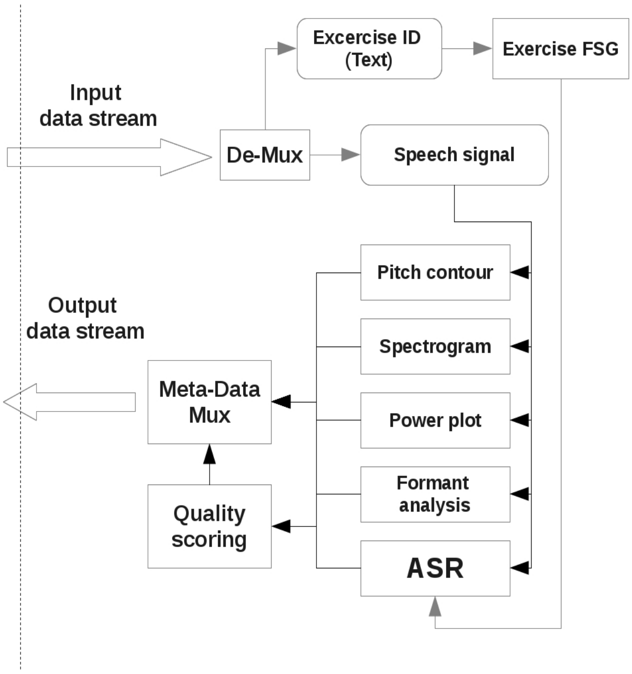


**Figure 1 -** System architecture

The phoneme segmentation component is based on the Pocketsphinx speech recognition engine [9] and its Gstreamer version. It is chosen because of the low processing and memory footprint: fast feedback to the user will be essential even when many clients connect at the same time and many instances of the engine might be running in parallel. The requirement real-time feedback 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 place LPC cepstrum is used for frame-based pitch determination, similarly LPC spectra peak picking method 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 segmentation 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 [10]. The 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 a separate finite-state grammar for each of the exercises containing a single state sequence without alternatives which corresponds to the words to be spoken.

## 4 Pronunciation quality scoring

The users' speech is compared against a reference speaker rather than to a (statistical) model being optimized on a large set of speakers. The intention is to resemble a training process conducted by a therapist, where the most important element is that the patient listens to the referent voice of his trainer and tries to imitate it as close as possible. For scoring of articulation, log-likelihood scores obtained from forced-alignment recognition with comparison against a reference utterance were used [11]. For the prosody (global and segmental) scoring, all four basic speech parameters: pitch, short time energy, spectrogram and formants, are first normalized to suppress the differences in the levels of intensity (different recording conditions) and the F0 frequency (different speakers) providing more consistent quality scoring [12].

The data was normalized for each utterance per each component (for spectrogram – number of frequency bins, for the formant contours – the $F_1$, $F_2$, ... $F_n$ frequencies). The values for minimum and maximum were chosen from the 95% confidence interval and the data values are transformed in the 0-1 range. The outliers smaller than 0 are set to 0, those larger than 1 are left as they are. For spectrogram the magnitude per frame and per frequency bin is normalized by the maximal amplitude in all frequency bins. Normalized measures were aligned and compared using Dynamic Time Warping (DTW) algorithm with the referent data, giving the RMSE (Root mean square error) along with the number of deletions (*D*) and insertions (*I*), normalized with the number of frames in the alignment path and the number of the components in the data structure. Phoneme and silence durations were determined from the phonetic labels derived by the forced-alignment recognition and normalized by the total duration of the speech (excluding silence periods). Acoustic (*ACS$_n$*) and duration (*Dp$_n$*) scores were estimated for each pair of user and reference phoneme with the following expressions:

$$ACS_n = 1 - \left| \frac{ACS_{ref} - ACS}{ACS_{ref}} \right| \qquad (1)$$

$$Dp_n = 1 - \left| \frac{Dp_{ref} - Dp}{Dp_{ref}} \right| \qquad (2)$$

The final *GOP* scores are composed of the relative difference between acoustic scores per phoneme (*ACS$_n$*) obtained with acoustic model trained on large speaker population, as well as the relative difference in the phoneme duration (*Dp$_n$*). These two scores are composed in general GOP score using a balancing factor (default $\alpha = 0.5$).

$$GOP = \alpha \cdot ACS_n + (1 - \alpha) \cdot Dp_n \qquad (3)$$

All contour matching measures and per phoneme GOP scores are separately transformed into z-scores (according to the statistical analysis on the available speech database) and presented back in appropriate form to the patients or the therapists.

## 5 Results and discussion

The proposed framework was used to analyze recordings from the collected database. The quality scores were estimated for all subjects (25 test and 3 control) on selected exercise set of 34 sentences from the diagnostic test chosen to address the quality of the articulation. Statistical analysis is required to discover the dependencies and the optimal contribution of the separate scores and to produce one score instead of relative ones. The general score should describe the overall quality including the articulation and the intonation goodness. Additional issue is the sensitivity of the phoneme segmentation to recordings which introduce significant mismatch with the used acoustic model. Environmental factors as the device, room, speaking distance and speakers with lower voice quality, could produce unexpected results. For example, speakers with breathy voice could introduce wrong segmentation of the front phoneme in the sentence due to notable energy presence, the resulting effect is that the segment duration is longer with lower acoustic likelihood. In general, the main requirement is that the speaker should pronounce the exercise as similar as possible with the referent utterance, therefore the choice of the referent speaker(s) is of highest importance.

### 5.1 Speech parameters contour matching

The contour matching results from 952 samples (sentences) were analyzed. Knowing the mean and the standard deviation, provides the possibility of transferring the score values in the same range where they can be transformed into z-scores. Various statistical tests revealed the interdependency of the analyzed parameters. Strong correlation between the *D* (deletions) was observed, as well as for the *I* (insertions) parameters, as well as for *I* and *D* across different parameters. However, the RMSE values are almost uncorrelated with the *D* and *I* measures, which suggest that RMSE score could be used in combination with them and not introducing redundancy into the scoring procedure. The Wilcoxon Mann-Whitney Test was used as a non-parametric test on the control (102) and the test group (850 sentences) to pinpoint which of the measures significantly differs between the speaker groups. The main hypothesis is that, “the control group achieves better result than the test group”, which means lower values in the control group for the proposed measures.

| Score | P-value (95%)<br>C-T <> 0 | P-value (95%)<br>C-T > 0 | Significance |
|---|---|---|---|
| **D1** | 0,9388 | 0,4694 | |
| **D2** | 0,0065 | 1,0000 | *** |
| **D3** | 0,6960 | 0,6521 | |
| **D4** | 0,0002 | 0,9999 | ** |
| **I1** | 0,0000 | 0,0000 | *** |
| **I2** | 0,0028 | 0,0014 | ** |
| **I3** | 0,0001 | 0,0000 | *** |
| **I4** | 0,0007 | 0,0003 | ** |
| **rmse1** | 0,0000 | 0,0000 | *** |
| **rmse2** | 0,0010 | 0,0005 | ** |
| **rmse3** | 0,0100 | 0,0950 | * |
| **rmse4** | 0,0007 | 0,0003 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

**Table 1** - Wilcoxon Mann-Whitney Test between the control (N=102) and the test group (N=850), Confidence 95%

The short time energy (1), spectrogram (2) and the formant track (4) contour matching significantly differ across the groups, while the pitch contour (3) matching, does not. The reason is that there is less difference in the F0 contours in the groups due to normalization process. Larger analysis frames were employed in the pitch estimation to cover at least two pitch periods, introducing smaller number of frames used in DTW alignment. Other important reason is, there are speakers with good intonation scores in the test group, but also speakers with bad intonation scores in the control group. This is a direct result of the recording procedure, which differs from the case where the patients should imitate the reference voice.

## 5.2 Phoneme articulation quality measures

The descriptive statistics of the GOP scores per each phoneme were calculated over 13484 occurrences (1461 control and 12023 test). The statistics were used to convert the relative scores to z-scores providing more meaningful user feedback and visualization. The phoneme distribution for the exercises shows that all the phonemes are appropriately represented which supports that the exercises content is phonetically balanced. The Wilcoxon Mann-Whitney Test between the groups confirms that there is a difference in articulation quality where the control group achieved better pronunciation scores than the test group, which was expected.

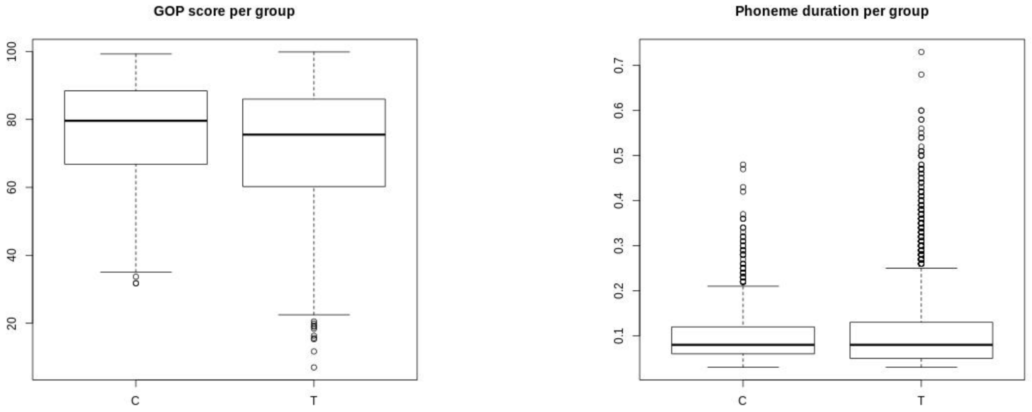
| Score | P-value (95%)<br>C <> T | P-value (95%)<br>C < T | P-value<br>(95%)<br>C > T | Significance |
|---|---|---|---|---|
| GOP | 0,0000 | 1,0000 | 0,0000 | *** |
| Duration | 0,1855 | 0,0928 | 0,9072 | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

**Table 2** - Wilcoxon Mann-Whitney Test for GOP scores and phoneme duration per group

The Figure 2 presents the comparison between groups of the GOP scores and phoneme durations. It can be seen that the control group achieves better results (higher mean, low deviation, less outliers) than the test group.

The speakers in the experiments are divided in test and control group only by the fact that they are healthy or subjects with speech disorder, the patient's diagnosis (level of disorder) is not taken into account. Apart from that, the speakers from the control group produced speech that is prosodicaly different compared to the referent speaker, as well as the speakers in the test group, because the recording procedure does not resemble the training procedure. In the recording sessions, the speakers were asked to read prepared sentences, not to imitate a reference speaker.



a

b

**Figure 2 -** GOP (a) and Phoneme duration (b) per group (C-control, T-test)

The reference speakers also could produce articulation and intonation errors, which have to be avoided by speech post-processing. The same statistical tests were performed including smaller number of test subjects, comparable with the size of the control group, and the results confirm the previous observations. In order to get more conclusive insight in the quality scoring process, each recorded sentence should be evaluated and scored by a human expert and the correlation with the automatic scores observed. Future research will be focused on quality scoring analysis with the speech data collected from the field tests.

## 6 Conclusions

In this paper, a Computer-Aided Speech Therapy system for dysarthric speakers based on a client-server platform is presented. The users perform audio-visually presented exercises, uttering words, word-pairs, phrases or longer texts, where the pronunciation and the prosody are compared against reference speech. The intention is to resemble a training process conducted by a therapist, where the patient listens to the reference voice of his trainer and tries to imitate it as close as possible. Learning progress is monitored for users and therapists in a users profile by exercise completion and success, as well as the improvement in pronunciation quality all observed over time. The client-server architecture allows usage of PCs or mobile clients (tablets, smartphones or PDAs) for remote access to the exercises. The recorded speech is streamed to the server in real-time for audio processing and the results are sent back to the client with the lowest possible latency. The results are presented in an adequate way, taking into account the user's cognitive capabilities and his assumed knowledge about speech, language and phonetics.

For testing and evaluation of the speech processing and scoring concepts, a domain-specific speech corpus was recorded. The recorded subjects are German native speakers divided into reference (2 healthy subjects), main (25 dysarthric subjects) and control (3 healthy subjects) group, recorded under similar technical conditions, in office or living room environment. For articulation scoring, comparison of phoneme log-likelihood scores against the reference speech data was used. Pitch, short time energy, spectrogram and formants were normalized and similarity measures compared against the reference speech were computed by DTW algorithm. The dependency and the optimal contribution of each measure to the overall score was statistically assessed over the available corpus. The Wilcoxon Mann-Whitney test showed that all measures (except pitch based) significantly differ across the speakers groups (healthy and dysarthric). The reason is, that in the recording procedure, the speakers were asked to read prepared sentences, not to imitate a reference speaker. For phoneme articulation scores, the same statistical tests confirmed that there is a significant difference in phoneme articulation across the groups. In general, the control group achieved better averaged pronunciation scores than the test group, which was expected. In order to get more conclusive results, each recorded sentence should be observed and evaluated by human experts to assess the correlation with the automatic estimated quality scores. Future activities will be focused on quality scoring analysis with the speech data collected from the field tests.

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

## References

[1] H. Strik, E. Sanders, M. Ruiter, and L. Beijer, “Automatic recognition of dutch dysarthric speech: a pilot study,” in Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '02), pp. 661–664, 2002.

[2] Maier, A., Haderlein, T., et al., "Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer", EURASIP Journal on Audio, Speech, and Music Processing, Article ID 926951, 2010.

[3] Donegan, M., (2000) Voice Recognition Technology in Education – Factors for Success, ACE Publications.

[4] Vaquero, C., Saz, O., Rodrıguez, W. R., & Lleida, E. (2008). Human Language Technologies for speech therapy in Spanish language. *Proceedings of the LangTech2008*, 129-132.

[5] Kitzing, P., Maier, A., & Åhlander, V. L. (2009). Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phonatrics Vocology*, *34*(2), 91-96.

[6] Jokisch, O., Koloska, U., Hirschfeld, D. and Hoffmann, R. “Pronunciation learning and foreign accent reduction by an audiovisual feedback system”. Proc. 1st Intern. Conf. on Affective Computing and Intelligent Interaction (ACII), Beijing, China, 2005, pp. 419-425

[7] 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.

[8] Fiukowski, H. (2004): Sprecherzieherisches Elementarbuch. Niemeyer, Tuebingen

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

[10] 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.

[11] Witt, S., and Young, St. (1997). Computer-assisted pronunciation teaching based on automatic speech recognition. Language Teaching and Language Technology Groningen, The Netherlands

[12] Flynn, N. (2011). Comparing vowel formant normalisation procedures. *York Pap. Linguist., Ser*, *2*(11), 1-28.