# SOFTWARE TOOLS FOR SPEECH THERAPY AND VOICE QUALITY MONITORING

*Oytun Turk and Levent M. Arslan*

Bogazici University Multimedia Laboratory (BUMM), Bebek, 34342, Istanbul, Turkey
R&D Dept., Sestek Inc., ARI-1 Teknopark Binasi, 34469, Istanbul, Turkey
phone: + (90 212) 286 25 44, fax: + (90 212) 286 25 47, email: <oytun, levent>@sestek.com.tr
web: www.bumm.boun.edu.tr www.sestek.com.tr

## ABSTRACT

This study focuses on the development of software tools integrated with speech processing technology for speech pathologists and for patients with speech/language disorders and voice quality problems. An integrated interface called CATSEAR is under development for database collection, data analysis, therapy design, and patient monitoring. Automatic assessment techniques using pattern recognition algorithms enable the speech pathologist to employ objective criteria during speech therapy as well as guide the patients when the pathologist is not available. The performance of the patient can be monitored over time. CATSEAR enables sharing of databases among speech therapists with remote collaboration and pre-recorded analysis facilities. It can also be used for relatively mild disorders like mispronunciation and for singing voice training.

## 1. INTRODUCTION

Speech and language disorders adversely affect the communication skills of at least 3.5% of the human population according to World Health Organization statistics. The degrees of disorders vary from mild impairments like pronunciation errors to more severe ones including hearing-loss, aphasia, and cranio-facial anomalies. Speech therapy aims to provide therapeutic assistance to individuals with all degrees of impairments.

The improvements in speech and signal processing technology have resulted in the development of computer-assisted methods for speech therapy [1], [2]. In [3], the assessment of articulation problems in children is formulated while in [4] monitoring in speech therapy is addressed.

The main objective of this study is the development of an interface called CATSEAR for speech therapists and patients that will be integrated with speech processing technology. The interface consists of analysis, recognition, and reporting modules. It will be useful as a remote monitoring tool for the therapist. Furthermore, it may serve as a replacement of the therapist when s/he is not available.

This paper presents the current state of the CATSEAR development project. Figure 1 shows the system flowchart of CATSEAR. Section 2 summarizes the overall system design and describes the modules. In Section 3, we discuss the results of preliminary tests for automatic speech recognition for articulation scoring. The paper is concluded with a discussion of the results and future work in Section 4.

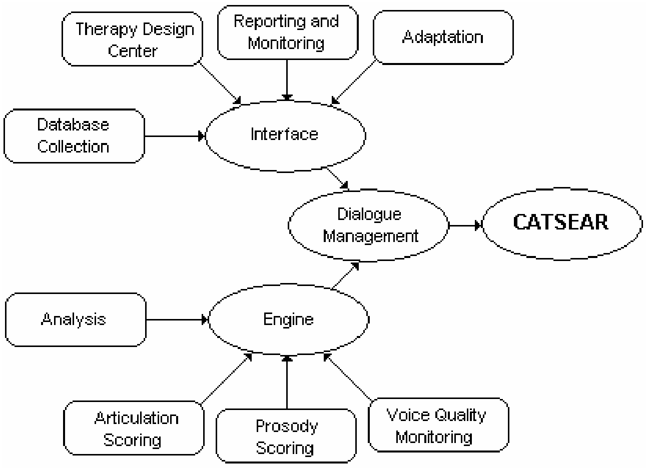


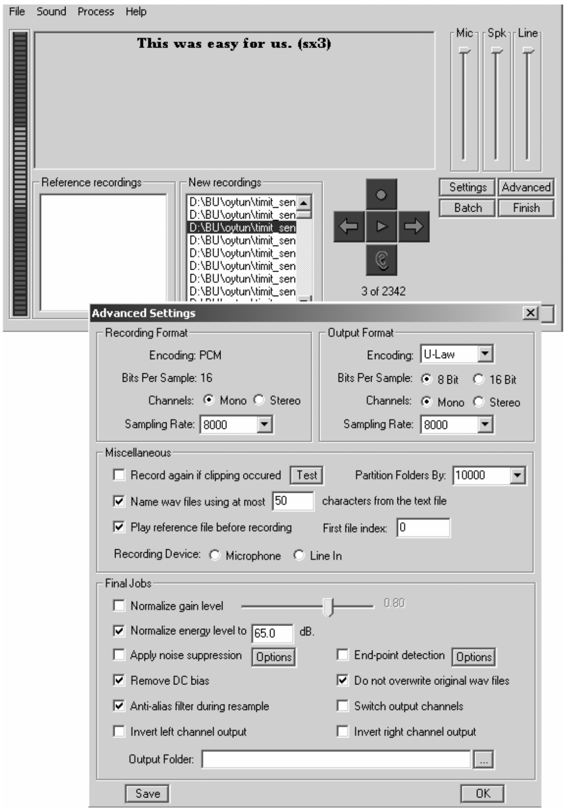
**Figure 1.** System flowchart for CATSEAR.

## 2. DESCRIPTION OF CATSEAR MODULES

Signal processing algorithms for CATSEAR are being developed in C/C++. The interfaces are designed using Microsoft Visual Studio 6.0 and .NET architectures. CATSEAR consists of three major modules: “Interface”, “Engine”, and “Dialogue Manager”. The “Interface” includes modules for database collection, therapy design, reporting, monitoring, and adaptation. The “Engine” provides feature analysis as well as automatic assessment through pattern recognition. The outputs of the “Interface” and the “Engine” are processed by the “Dialogue Manager” to create a highly interactive and user-friendly interface for the therapist and the patient. We briefly describe the individual modules in the following sub-sections.

### 2.1. Database Collection Module (DCM)

DCM supports collection of databases of (i) speech-only signals, (ii) speech and electro-glottograph (EGG) signals, and (iii) speech and video signals along with 3-D markers. Figure 2 shows a snapshot of DCM.



This was easy for us. (sx3)

Reference recordings

New recordings
D:\BU\oytun\timit_sen
D:\BU\oytun\timit_sen
D:\BU\oytun\timit_sen
D:\BU\oytun\timit_sen
D:\BU\oytun\timit_sen
D:\BU\oytun\timit_sen
D:\BU\oytun\timit_sen
D:\BU\oytun\timit_sen

3 of 2342

Advanced Settings

Recording Format
Encoding: PCM
Bits Per Sample: 16
Channels: ● Mono ○ Stereo
Sampling Rate: 8000

Output Format
Encoding: U-Law
Bits Per Sample: ● 8 Bit ○ 16 Bit
Channels: ● Mono ○ Stereo
Sampling Rate: 8000

Miscellaneous
□ Record again if clipping occured [Test]
Partition Folders By: 10000
☑ Name wav files using at most 50 characters from the text file
☑ Play reference file before recording
First file index: 0
Recording Device: ○ Microphone ○ Line In

Final Jobs
□ Normalize gain level 0.80
☑ Normalize energy level to 65.0 dB.
□ Apply noise suppression [Options]
□ End-point detection [Options]
☑ Remove DC bias
☑ Do not overwrite original wav files
☑ Anti-alias filter during resample
□ Switch output channels
□ Invert left channel output
□ Invert right channel output
Output Folder:

**Figure 2.** Database collection module.

### 2.2. Signal Analysis Module (SAM)

SAM consists of standard algorithms for speech, EGG, and image analysis. It extracts temporal and spectral features as well as phonetic information for the automatic assessment modules. It is integrated with an HMM-based speech recognizer and an interface for visualization. Figure 3 shows the visualization interface.

### 2.3. Prosody Scoring Module (PSM)

PSM uses the pitch and energy contour of a given utterance along with its phonetic transcription and compares them with reference contours to generate a pitch-match and an energy-match score. The scores are computed by aligning the given and the reference utterances using the phonetic transcription information and computing the normalized cross-correlation coefficient between the pitch contours and energy contours of the given and reference utterances. An example is shown in Figure 4.

### 2.4. Articulation Scoring Module (ASM)

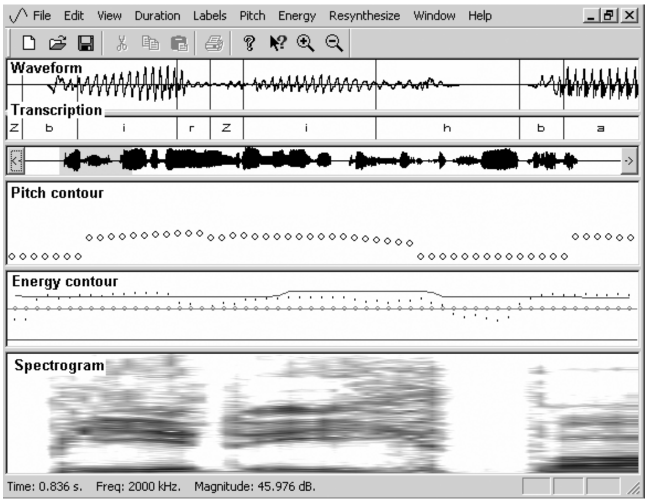
ASM performs automatic articulation scoring based on the information obtained from SAM. It provides phone-level, word-level, and sentence-level assessments in order to provide a detailed description of the articulation performance.

### 2.5. Voice Quality Monitoring Interface (VQMI)

VQMI extracts different voice quality parameters including high-to-low harmonic energy ratio, jitter, and shimmer in real-time to provide feedback using OpenGL based computer graphics as shown in Figure 5. It supports MIDI controllers that can be used for musical input in voice quality exercises and singing voice training.

### 2.6. Therapy Design Center (TDC)

TDC supports the design and application of different speech therapy exercises. It provides facilities to create, edit, and save the exercises as well as remote-sharing for online access of the patients and remote collaboration among speech therapists. Figure 6 shows a snapshot from the beta version of TDC.



Time: 0.836 s. Freq: 2000 kHz. Magnitude: 45.976 dB.

**Figure 3.** SAM visualization interface.

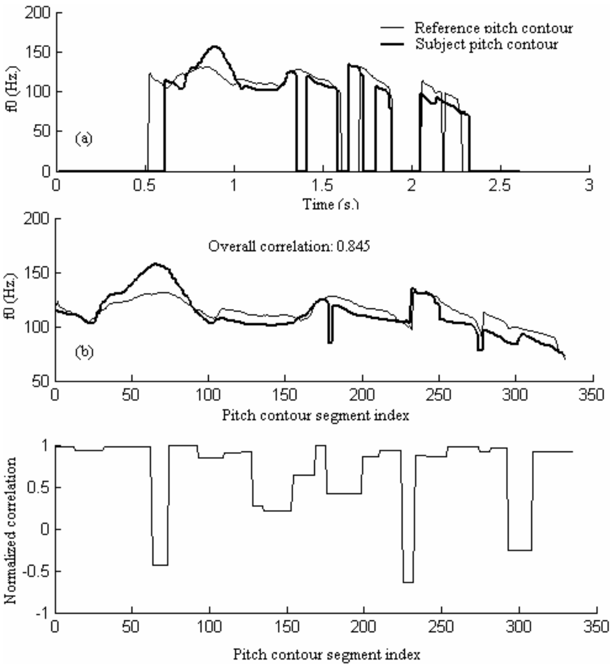
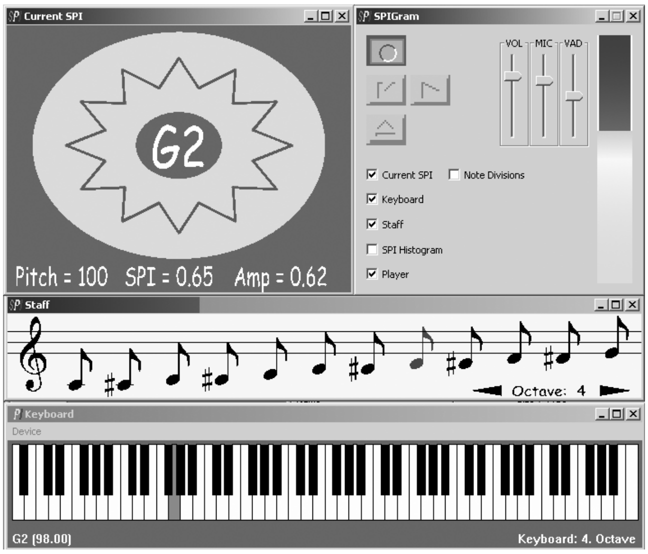


**Figure 4.** (a) Reference and subject pitch contours to be compared corresponding to the utterance in Turkish “Kaza nedeniyle ulaşım aksadı” from two different speakers. (b) Time-aligned pitch contour profiles (TA-PCP) for the reference and subject pitch contours showing the corresponding f0 values in the two contours. (c) Segmental scores (i.e. the normalized correlation coefficient between corresponding

pitch contour segments) that show the match between the subject and reference pitch contours.



Pitch = 100 SPI = 0.65 Amp = 0.62

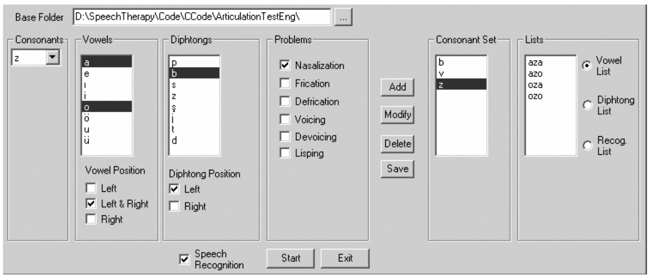
☑ Current SPI □ Note Divisions
☑ Keyboard
☑ Staff
□ SPI Histogram
☑ Player

G2 (98.00) Keyboard: 4. Octave

**Figure 5.** Snapshot of the voice quality monitoring interface.

### 2.6. Therapy Design Center (TDC)

TDC supports the design and application of different speech therapy exercises. It provides facilities to create, edit, and save the exercises as well as remote-sharing for online access of the patients and remote collaboration among speech therapists. Figure 6 shows a snapshot from the beta version of TDC.



Base Folder D:\SpeechTherapy\Code\CCode\ArticulationTestEng\

Consonants: z

Vowels: a e ı i o ö u ü

Vowel Position: □ Left ☑ Left & Right □ Right

Diphtongs: p b s z ş j t d

Diphtong Position: ☑ Left □ Right

Problems: ☑ Nasalization □ Frication □ Defrication □ Voicing □ Devoicing □ Lisping

Consonant Set: b v z

Lists: aza azo oza ozo

● Vowel List ○ Diphtong List ○ Recog. List

☑ Speech Recognition

**Figure 6.** Therapy design center (beta version)

### 2.7. Dialogue Management Module (DMM)

DMM will interact with the user using text-to-speech synthesis, 3-D face synthesis, and animated agents to provide a highly interactive tutoring environment even in the patient's home. It will also help to attract children who constitute a major portion of a speech therapists client profile. DMM is currently under development. Figure 7 shows two synthetic faces available in DMM.

### 2.8. Reporting and Monitoring Module (RMM)

RMM will provide an integrated interface for progress monitoring in the speech therapy process. Speech therapists will have access to long-term data as well as statistics on the patient's performance. RMM is currently under development.

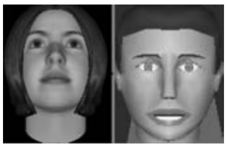


**Figure 7.** Speech driven 3-D facial animation (beta version).

### 2.9. Adaptation Module (AM)

AM, which is currently under development, will provide feedback to supplement the speech therapist and/or to automate the therapy process. It will consist of re-trainable pattern recognition algorithms for adaptation to different learner's situations including Hidden Markov Models for speech recognition, Artificial Neural Networks for automatic diagnosis, and classification algorithms (K-Means, Self Organizing Maps, LGB, etc.). AM will be integrated with speaker independent models which are trained using large databases and will also support extension with new features.

## 3. EVALUATIONS

In a preliminary test, we have evaluated the performance of a standard HMM based speech recognizer [5] for isolated word recognition between pairs of confusable word pairs in Turkish. The list of word pairs were obtained from a group of speech therapists who used them in articulation tests. The words in each pair differed only in one consonant. The following list of confusable consonants is used: /f/-/v/, /s/-/z/, /c/-/ç/, /k/-/g/, /r/-/y/, /ş/-/j/, /t/-/d/ and /p/-/b/.

| Consonant pairs | Start of word | Middle | End of word |
|---|---|---|---|
| /f/-/v/ | fidan-vidan | defter-devter | çarşaf-çarşav |
| /s/-/z/ | sarı-zarı | asker-azker | kas-kaz |
| /ş/-/j/ | şale-jale | ajan-aşan | beş-bej |
| /c/-/ç/ | ceket-çeket | acı-açı | avuç-avuc |
| /p/-/b/ | pasta-basta | kapı-kabı | dolap-dolab |
| /t/-/d/ | tabak-dabak | katı-kadı | yakut-yakud |
| /k/-/g/ | kez-gez | basket-basget | gözlük-gözlüg |
| /r/-/y/ | raket-yaket | çorap-çoyap | bir-biy |

**Table 1.** List of confusable word pairs generated by substitution of the confusable consonant pairs.

The words given in Table 1 are recorded from 10 speakers and the recognition performance is measured using 10-fold cross validation. For each cross-validation step, the recordings of one speaker were reserved as the validation set and the HMMs were trained using the rest of the database. The recognition performance is measured as the average of all cross-validation steps. The average rate was 80.3% for the list of confusable word pairs given in Table 1. However, further tests are required for detailed performance evaluation of speech recognition based articulation scoring.

## 4. CONCLUSIONS

An integrated interface called CATSEAR is under development for speech therapy and voice quality monitoring. It will assist the therapists in database collection, data analysis, therapy design, and patient monitoring. The patients will be able to exercise when the therapist is not present and monitor the improvements in their skills with the automatic assessment tools.

As future work, we will collect databases of different speech and language disorders and test the pattern recognition algorithms. The voice quality monitoring interface will be standardized by collecting statistics of different voice quality parameters. All modules will be integrated in a single executable program. We are also planning to develop a test framework for CATSEAR in order to evaluate its performance in real-life scenarios.

## 5. ACKNOWLEDGEMENTS

Part of this research is supported by the European Union Network of Excellence SIMILAR (IST-507609) and Tübitak, Turkey (TIDEB 3040039).

## REFERENCES

[1] Witt, S. M., and Young, S. J., "Phone-level pronunciation scoring and assessment for interactive language learning", in Speech Communication*,* 30 (2-3), pp. 95-108, (2000).

[2] Herron, D., Menzel, W., Atwell, E., Bisiani, R., Daneluzzi, F., Morton, R., and Schmidt, J. A. "Automatic localization and diagnosis of pronunciation errors for second-language learners of English", in Proc. of the Eurospeech 1999, pp. 855-858.

[3] Bunnell, H. T., Debra, M. Y., and Polikoff, J. B., "Using Markov models to assess articulation errors in young children", in The Journal Of The Acoustical Society of America, Vol. 107, Issue 5, p. 2093, (2000).

[4] Bunnell, H. T., Yarrington, D. M., and Polikoff, J. B., "STAR: Articulation training for young children", in Proc. of the ICSLP 2000, Vol. 4, pp. 85-88.

[5] Rabiner, L. R., and Juang, B.-H., Fundamentals of Speech Recognition*,* Prentice-Hall, Inc., New Jersey, NJ, 1993.