A SYSTEM FOR TEACHING SPEECH TO PROFOUNDLY DEAF CHILDREN USING SYNTHESIZED ACOUSTIC & ARTICULATORY PATTERNS

Elizabeth Keate, Hector Javkin, Norma Antonanzas-Barroso, Ranjun Zou
Speech Technology Laboratory, Panasonic Technologies Inc.,
Santa Barbara CA, 93105 U.S.A .

ABSTRACT

This paper describes a computer assisted method of teaching profoundly deaf children to speak, which employs the unique feature of an integrated text-to-speech system (TTS). Our earlier speech training system [1] presented a series of speech parameters, derived from articulatory instruments and acoustic analysis, in a visual form. In that system, teacher's speech is input to the system and used as a model for the children to follow, and the children's speech is monitored to provide feedback. As with other computer-aided speech training systems (e.g. [2]), the teacher-assisted trainer is limited by the time students have with speech teachers. Several computer-based systems for providing information as to the desired acoustic and articulatory patterns and feedback showing what the children are doing already exist. In our system, we have developed an articulatory component which synthesizes tongue-palate contact patterns for the children to follow.

1. INTRODUCTION

Teaching profoundly deaf children to speak is a difficult challenge. The deaf who desire to speak have to learn articulatory and acoustic gestures to produce speech without the benefit of acoustic feedback.

Even with the technological advances in recent years of computer-based speech training systems for deaf children [2,3], the systems are still limited by the need for a teacher to provide the speech production models. Typically, the time that children have with teachers is restricted, and without a teacher's assistance, children can practice only with pre-stored utterances. This paper describes a method which integrates a text-to-speech system [4] to synthesize production models, including tongue-palate contact patterns, to provide children with an enhanced speech training environment. As with any full text-to-speech system, the number of utterances that can be generated are infinite, thus providing children with the independence to study on their own.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association of Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
ASSETS 94- 10/94 Marina Del Rey, CA, USA
© 1994 ACM 0-89791-649-2/94/0010..$3.50

2. LIMITATIONS OF EXCLUSIVELY-ACOUSTIC OR ARTICULATORY BASED SPEECH TRAINING SYSTEMS

Conceivably, if a deaf person is provided sufficient visual information as to the articulatory and acoustics of speech, they could have the same capability for learning speech as a hearing person. However, even if the visual representation of the acoustic and articulatory signals were complete, a problem persists. There is a vast difference between the amount of time that hearing persons have access to acoustic information and the amount of time deaf persons have access to a visual representation of the acoustics. As a result, a deaf child requires instruction that is extremely focused in providing the speech information efficiently.

2.1. ACOUSTIC TRAINING

One method to assist in acoustic training is by using a dynamic palatograph. The dynamic palotagraph, used in our system, is illustrated in Figure 1. The palatograph [2,3] provides contact data via an artificial palate mounted with a series of electrodes which pass a low-amperage current whenever they are touched by the tongue.

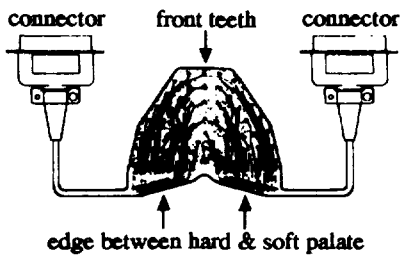


Figure 1. *System Configuration of Dynamic Palatograph.*

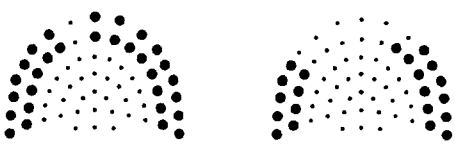


Figure 2. *Tongue Contact for a Hearing Speaker (left) Saying /s/. Speaker Approximating Contact for /s/ (right)*

Figure 2 above (left) shows the contact pattern on the palate for a hearing speaker saying /s/, while figure 2 (right) shows the contact pattern on the palate for a speaker who is attempting to say /s/ but has not quite learned the appropriate tongue-palate contact pattern. An acoustic-based system cannot recognize that this is a relatively close approximation to the desired contact, because there is insufficient constriction to cause frication noise, so that the acoustic realization is totally different from /s/. Even a labiodental fricative such as /f/ will produce a closer acoustic result, and so that a deaf speaker, relying solely on a visual representation of the acoustic signal, will be led away from the correct production. Hearing persons usually learn to produce /s/ despite this difficulty, but they have acoustic feedback available full-time. Even for hearing speakers, learning speech on the basis of the acoustic signal is often imperfect, leading to changes in the way that languages are pronounced [6,7,8]. Since deaf children receive a visual representation of the acoustic signal only at training sessions, they can derive substantial benefits, in learning to speak, from receiving articulatory information and observing their own successive approximations to the correct articulations.

2.2. ARTICULATORY TRAINING

The sound /s/ also reveals one of the difficulties with the purely articulatory approach, because the exact articulation varies from speaker to speaker [9], largely depending on the individual physiological configuration of the teeth, jaw and tongue. The proper tongue position for a particular configuration is difficult to predict. Therefore, even after a close approximation of an /s/ articulation is achieved, a deaf child requires feedback about the acoustic signal to articulate an /s/ in the most suitable way for his or her mouth.

Vowel sounds, which are produced by resonances created in the vocal tract, are also very difficult to train on the basis of articulatory position. Furthermore, much of the articulatory pattern, such as the constriction of the pharynx, cannot be determined without relatively invasive techniques. However, since the frequency locations of the formants, especially F1 and F2, are closely associated with the shape of the vocal tract as the lips, tongue, pharynx, and jaw move to articulate the vowel. Deriving information from analyzing the frequency space of the formants F1 and F2 can provide sufficient information to be used in teaching the production of vowels.

3. SPEECH TRAINING SYSTEM OVERVIEW

This speech training system is a integrated PC based device which uses a modified STLtalk synthesizer [4] to provide speech model input as well as speech output, and an acoustic and artriculatory sensor system which extracts the training parameters from the student.

3.1. INPUT SENSOR SYSTEM

The input sensor system is based on the CISTA system developed by members of our group and colleagues [5]. The elements of the input sensor system and how they are worn can be seen in figure 3.

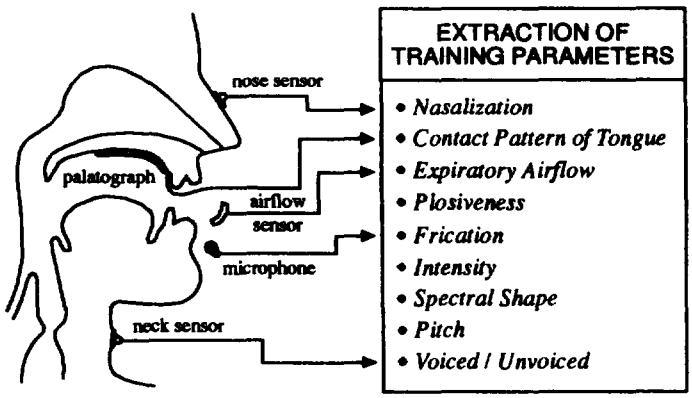


Figure 3. *Diagram of Sensor Instruments as Worn, and Associated Data Collected*

The dynamic palatograph already described provides tongue-palate contact data. An accelerometer pressed lightly to the nose provides a measure of nasality. An accelerometer on a neck collar provides voicing on-off information even with very weak or breathy voicing. An airflow meter hand-held in front of the face provides qualitative airflow information. The acoustic signal, used for amplitude, fundamental frequency, and spectral shape, is provided by a headset-mounted microphone. The system is designed to provide integrated articulatory and acoustic data in both a clinician-oriented form and in the form of video games.

3.2. STLTalk TEXT-TO-SPEECH SYSTEM

The STLTalk TTS system [4] is the synthesizer used in our speech training system. This system inputs raw ascii text via the RS-232 interface. This data is then segmented and processed into four data structures. These structures are then sent to the phonetic module which calculates phoneme durations, stress, and intonation for the text segment and generates 21 control parameters every 10 msec, are input to the formant synthesizer. The formant filter structure consists of five second-order IIR filters for voiced sounds, a pole-zero pair for nasal consonants, and a parallel filter structure of five second order IIR filters for fricatives. Output from the formant synthesizer is sent to a D/A converter and output speech is produced.

3.3. GRAPHICAL USER INTERFACE

In designing a interface for young children, simplicity is paramount. This system is designed with a window based graphical user interface which takes input from the keyboard or mouse. One of the goals of our system is

to allow the students to work independently of a teacher. However, children of this age do not have very good manual dexterity. Therefore, we developed a graphical, mouse-based interface with limited options and with icons that have almost four times the area of the icons in typical commercial systems designed for adults and older children. Some informal testing showed that children as young as 3 years could use mouse-driven computer programs. The screen image of the graphical user interface is shown in figure 4.

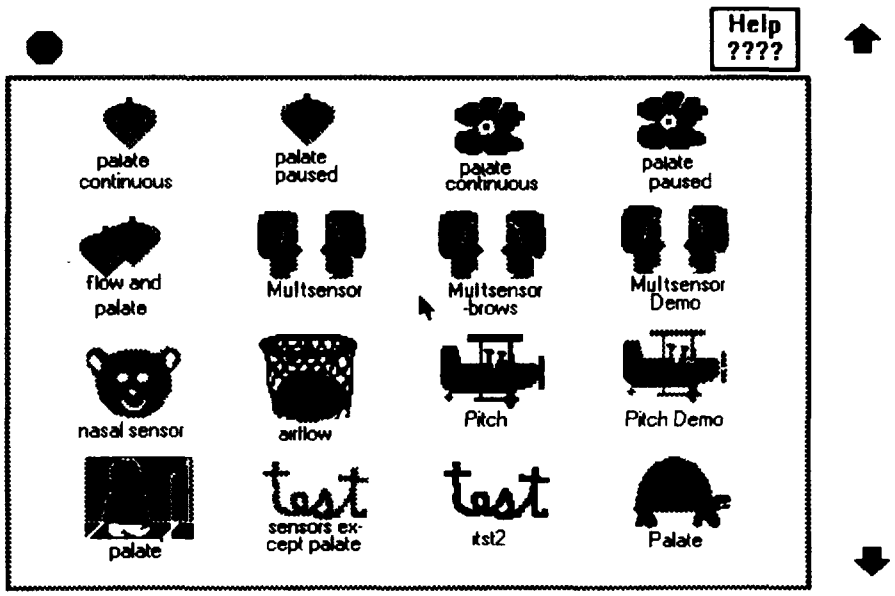


Figure 4. *Screen Image of the Graphical User Interface: Main Control Panel*

3.4. MOTIVATIONAL VIDEO GAMES

Given the age of the children, for which our system is intended, it is important to make the training system as interesting and as enjoyable as possible. Systems developed by BBN [2], IBM, Video Voice and others have provided feedback in the form of video games. These have been successful in motivating students to use the speech training systems. Our system is currently composed of thirteen different speech teaching computer programs.

3.4.1. TONGUE-PALATE CONTACT

Four distinct games were designed to teach tongue-palate contact on the basis of palatographic data, one of which will be described here. The computer game is designed simply to make the standard palatographic display more attractive. The other three represent the tongue-palate contact as games that are less closely tied to traditional displays.

In this program, shown in figure 5, the palatographic display is shown on the back of a turtle. The legs of the turtle and the background move, creating the illusion of walking.

3.4.2. MULTI-PARAMETER PROGRAM

Given the importance of coordination of the different gestures in speech, a program was developed that simultaneously provides feedback for up to four different speech parameters: amplitude, nasalization, pitch, and the absence or presence of voicing. The program displays two people looking at each other, so that the screen shows a lateral view of each. Fundamental frequency or pitch is given by eyebrows that go up and down, amplitude is given by a mouth that changes size, nasalization is given by a nose that changes size, and the Adam's apple expands to indicate voicing.

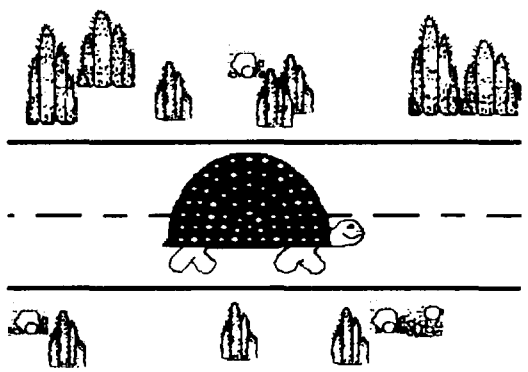


Figure 5. *Game Version of Palatographic Display*

3.4.3. PITCH CONTOUR PROGRAM

The multiple parameter program provides pitch feedback in a way that facilitates seeing pitch in relation to other parameters. However, it does not provide a pitch contour for the child to follow. The pitch contour program provides a dynamic contour for teaching variations in pitch.

3.4.4. AIRFLOW (STOP-BURST) PROGRAM

The stop bursts resulting from stop consonants lend themselves very well to games in which the initial velocity controls the outcome. The first derivative of airflow is a very good indication of a stop burst, as a proper burst requires a sharp increase in flow. We developed a video program to teach stop bursts.

4. TTS ASSISTED TRAINING SYSTEM METHOD

This TTS assisted method [5] creates model speech training parameters for any utterance the child types into a computer, and contains the basis for evaluating the child's speech production parameters against the model parameters. The modified text-to-speech system outputs, in addition to synthesized speech, the set of parameters which control the synthesis. This information is used to create parameters for modeling articulation for children. Part of the synthesized information does not represent articulation or acoustics, but rather what the output of instruments measuring articulation would be. The method is illustrated in Figure 6.

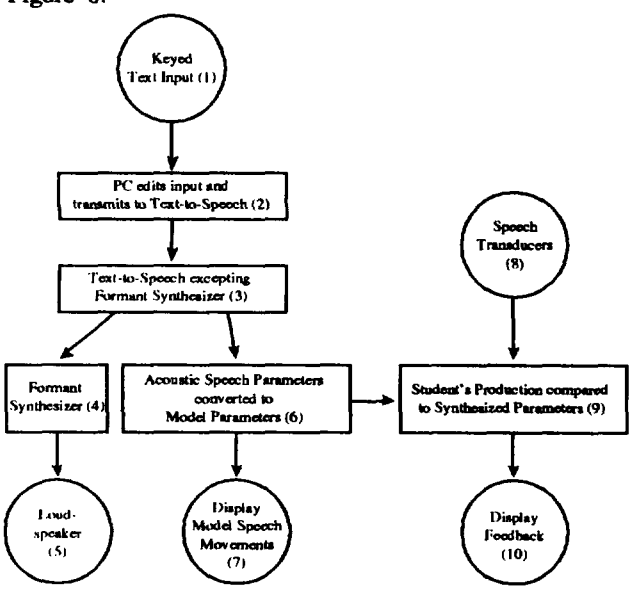


Figure 6. *Using Parameters Derived from TTS System for Speech Training*

The process is the following:

1. Using a computer keyboard, a student types in the utterance to be learned.
2. The utterance is reformatted and transmitted to the TTS system.
3. The text string is processed by the TTS system which produces phonetic parameters as well as phoneme labels and corresponding durations.
4. The parameters are input to a formant synthesizer to produce a 10kHz sampled digital speech waveform which is output to a digital analog converter.
5. A loudspeaker makes the acoustic signal available for the teacher's monitoring and for those students who have some residual hearing.
6. Concurrently with the speech output, the control parameters as well as the transition type and timing information for the second formant are sent to the processing unit for conversion to a form which can be used in teaching articulation.
7. The model speech movements for the student to imitate are displayed on a monitor.
8. The student's speech is monitored with the speech transducers described earlier.
9. The student's output is compared to the synthesized parameters. Currently, this part of the process is not automated, so that the child makes the comparison.
10. The student's output is displayed on a monitor for comparison with the synthesized output.

5. SYNTHESIS OF ARTICULATORY AND ACOUSTIC MODEL PARAMETERS

The aim of synthesizing the model parameters is to provide parameters equivalent to those that the training system measures and provides as feedback for the child. The synthesis of the model parameters differs depending on the particular parameter to be synthesized. Parameters such as fundamental frequency (F0), amplitude, and formant frequencies can be passed virtually unchanged from a text-to-speech system that synthesizes a child's voice. Other parameters must be synthesized or derived from the output of the synthesizer.

5.1. SYNTHESIS OF TONGUE-PALATE CONTACT PATTERNS

One of the important elements of our system is providing tongue-palate contact data for consonants such as /t/, /s/, /sh/, using a system of dynamic palatography [7,8,9]. To provide a model of the proper tongue-palate contact, the palatographic patterns have to be synthesized. TTS is used to provide the timing of four points: the onset of a sound, the time when maximum amplitude is achieved, the time when amplitude begins to decay, and the offset. These times are then coordinated to a set of palatographic images. The maximum area of contact for each sound is stored. Since speech sounds vary in contact pattern for different contexts, a number of different contact patterns need to be stored for each context, as well as for each

sound. The transitions between the contact areas for different sounds are governed by a set of rules.

Figure 7 shows one frame of the dynamic display a child sees on the screen. The palate contact shown on the right is that of the sound /z/ synthesized by the system. The display on the left is waiting for the child to produce her or his own contact pattern. The frame rate is currently being modified to 60 Hz, in order to be compatible with the screen frame rate of most computer monitors.

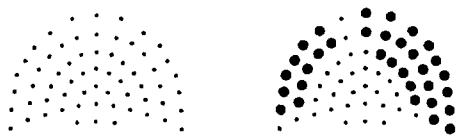


Figure 7. *Palatographic Pattern Used in Speech Training*

The following figure shows the utterance "This is it" as spoken by a native American English speaker. The frame rate of the contact measurement program is 100 Hz, which creates an offset when measured data is compared to synthesized data, as shown in the figure 8.

Figure 9 shows a sequential display of the contact pattern for the sentence "This is it" as synthesized by the system. Note that tongue-palate contact is considerable during the /IH/ vowel. Such contact is normally observed for so-called front vowels with our system. Methods of palatography, which contain electrodes going into the soft palate region, also recover contact for so-called back vowels.

A significant difficulty is that optimal contact patterns depend on each individual's configuration of the palate and teeth. This problem is partially solved by storing palates of good productions produced by the student. Children sometimes produce single instances of good articulations that they have difficulty repeating; by storing their productions, and using articulations selected by teachers as the models, they can receive instruction based on their individual tongue-palate contact configurations.

5.1.1. DISCUSSION

In summary, the benefits of providing synthesized articulatory and acoustic model data for the children to follow are twofold: 1) Children are able to practice while a teacher is not available, potentially greatly extending the time they can undergo effective training. 2) The synthesized models, unlike models that are pre-stored or produced by a teacher, can be adapted for each particular child, on the basis of their teacher's evaluation as to when children are approximating proper production. Although testing of the synthesized model component with deaf children is only now getting underway, we are hopeful that it will provide significant benefits in training. Computer-based training systems have been tested [9] for deaf children learning to

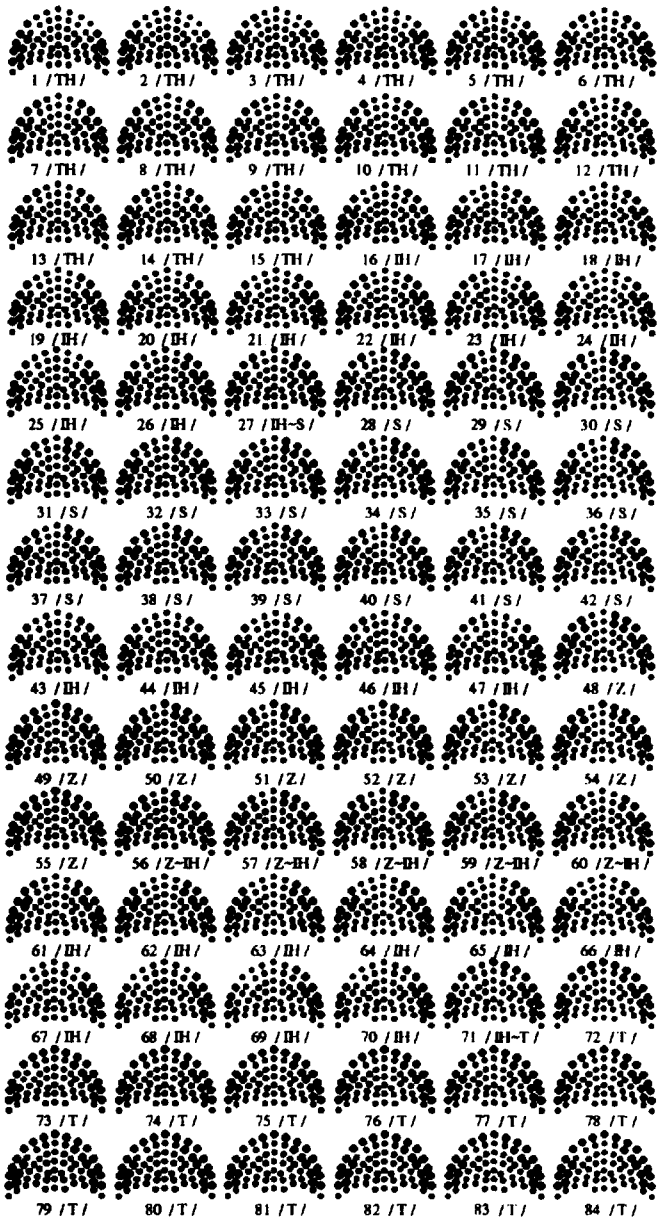


Figure 8. *Contact Pattern for a Speaker Saying "This is it."*

speak Japanese and have shown to be effective in improving the articulation of individual phonemes. The capability of a synthesized system of providing model information for words and sentences can be expected to offer additional benefits in the teaching of the articulation necessary for connected speech.

ACKNOWLEDGMENTS

We wish to gratefully acknowledge the advice, comments and assistance we received from Ted Applebaum, Edward Cudahy, Brian Hanson, Kazue Hata, David LaDelfa, Harry

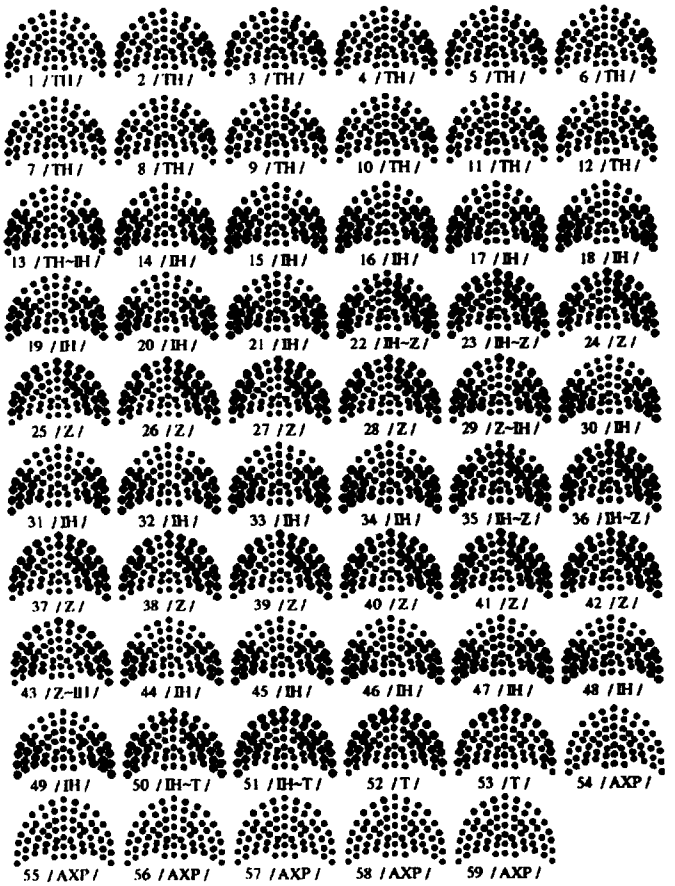


Figure 9. *Synthesized Contact Pattern for the Utterance "This is it."*

Levitt, Kenji Matsui, Tsuyoshi Mekata, N. Murata, Nancy Niedzielski, Eiichi Nonomura, John Ohala, Kristin Precoda and Hisashi Wakita.

REFERENCES

[1] Javkin, H., Antonanzas-Barroso N., Das A., Zerkle D., Yamada Y., Murata N., Levitt H., and Youdelman K., "A Motivation-Sustaining Articulatory/Acoustic Speech Training System for Profoundly Deaf Children," *Proc. ICASSP-93*, Volume 1, pp. 145-148, 1993.

[2] Potter, R. K., G. Kopp & H. Green, *Visible Speech*, New York, 1947.

[3] Kuzmin, Y. Mobile palatography as a tool for the acoustic study of speech sounds. Report of the 4th International Congress on Acoustics, Copenhagen. Paper 635, 1962.

[4] Javkin, H., K. Hata, L. Mendes, S. Pearson, H. Ikuta, A. Kaun, G. DeHaan, A. Jackson B. Zimmermann, T. Wise, C. Henton, M. Gow K. Matsui, N. Hara, M. Kitano, Der-Hwa Lin, Chun-Hong Lin, "A Multilingual text-to-speech system," *Proc. ICASSP 89*, pp. 242-245, 1989.

[5] Javkin, H., Keate, E., Antonanzas-Barroso, N., Yamada, Y., Youdelman, K. , "Automatic model parameter generation for the speech training of deaf children" Proc. ICASSP-94, Volume II, pp. 229-232, 1994.

[6] Allen, J., S. Hunnicutt, and D.H. Klatt, *From Text to Speech: The MITalk System*, Cambridge University Press, Cambridge, U.K., 1987.

[7] Fujimura, O., I. Tatsumi & R. Kagaya. Computational processing of palatographic patterns. *J. of Phonetics* 1:47-54, 1973.

[8] Nickerson, R.S., D.N. Kalikow and K.N. Stevens, Computer-aided speech training for the deaf. *J. of Speech & Hearing Disorders* 41:120-132. 1976.

[9] Yamada, Y. and N. Murata, "Computer Integrated Speech Training Aid," International Symposium on Speech and Hearing Sciences, Osaka, Japan, July 1991.

[10] Andersen, H. Abductive and Deductive Change. *Language* 49:765-793, 1973.

[11] Ohala, J.J. Experimental Historical Phonology, in J.M. Anderson and C. Jones (eds.) *Theory and Description in Phonology*. Amsterdam: North Holland Publishing Co. 353-389, 1974.

[12] Hardcastle, W. Dynamic Palatography. Work in Progress No. 2. Department of Phonetics and Linguistics. Edinburgh University, pp. 53-59. 1968.