A MODEL FOR AUTOMATED SPEECH CORRECTION OF GERMAN VOWELS: A PILOT STUDY

Rudolf Weiss

Western Washington University
Bellingham, WA, USA

Antonio Arroyo

University of Florida
Gainesville, FL, USA

ABSTRACT

A model is provided by which automatic error detection of vowels could be accomplished using predictable and pedagogically pretermined environments and specific analysis routines.

1. INTRODUCTION

Despite advances in modern technology and computerized speech recognition, a device which automatically detects and provides for correction of pronunciation errors is still far in the distance. However, we believe it is possible with the application of basic phonetic principles and pedagogical techniques to create an automated device which may be of use to foreign language teachers and students.

Criteria for automated speech recognition were already discussed nearly twenty years ago [1]. For the last ten years there has been almost a preoccupation with automatic segmentation and labelling of speech sounds, as can be seen by the large number of papers in this area at the last congress. It is exactly in this area of sound segmentation where some of the greatest problems in contemporary phonetics lie. With the advent of digitizing techniques, great strides have been made in voice analysis and synthesis. However, it has not changed the speech act itself. The problem, as has been pointed out for decades and again quite recently, is that of determining segment boundaries [3]. The efforts to find discrete information equivalent to sounds in the myriad of signals emitted by the continuous speech act eludes phoneticians. Ever present elements of co-articulation, compounded by factors of individual production and physiology, as well as suprasegmental elements and changing temporal aspects manifested in the continuous speech signal, confound efforts to find "sounds" particularly in an automated and error-free fashion. Furthermore it has been demonstrated that even we humans have difficulty labelling sounds categorically (absolutely) but must instead rely upon the contrastive relationship of the environment [2].

It is therefore our firm belief that at this point in the development of technology, error detection/correction can more easily be successfully accomplished if efforts are goal-directed to specific predictable errors and if they can be pedagogically "framed."

2. TECHNIQUES

A variety of techniques have been used in speech recognition, particularly for segment labelling.

Basically the incoming acoustic signal has to be broken down at certain intervals and then matched to some preprocessed acoustic criteria. The techniques vary; often a step filter device is used, matching bands of spectral energy above or below a certain frequency to the existence of certain sounds [6,3]. Routines of this nature often resemble a Jacobsonian distinctive feature approach.

3. WEISS MODEL

The model which we propose attempts to avoid some of the pitfalls inherent in conventional speech recognition processing. In application to foreign language teaching (or speech correction), certain assumptions have to be made:

- There are spectrographically identifiable and definable segments (sounds) in natural speech. We will create predictable environments for the ease of processing these targeted sounds.
- There are always features of co-articulation (transitions, alterations, etc.) in natural speech. These are also largely definable and predictable by the environment. By predetermining the environment we will be able to circumvent most of the problems co-articulation features might present.
- There are acoustic characteristics as well as idiosyncratic articulatory habits of each speaker. These are less predictable and no model can totally accommodate them.

A computerized model, using a Mac II and MacSpeech Lab II, or a comparable speech work station would work in the following manner:

- A correct utterance produced by a native speaker has been digitized and stored. The student is given a screen prompt and the digitized utterance is played.
- A prompt appears on the screen to repeat the word.
- The student's response is then digitized.
- The computer processes a hierarchy of matching routines (based on matching the digitized information).
- If an error is made, the student is prompted as to the nature of the error. The correct utterance is given again, and the student is prompted to repeat the utterance.
- The computer reprocesses, matches and gives error statements until the student responds correctly or gives up.

There are two real limitations to the functional success of such a model.

- The processing and computer response (digitizing, analyzing and matching routines) must be very fast (ideally < 1.5 sec) to be of practical use. Otherwise the nature of the student's production is likely forgotten.
- The model will work only if the errors are predefined and predictable and if the environment is completely controlled and chosen to facilitate ease of computer processing.

This model is an outgrowth of previous work on computer assisted diagnosis of vowel perception and a phonetics manual written by the author in which specific anticipated errors and exercises for overcoming these errors are provided for each sound [4,5]. This contrastive and predictive approach can be applied to our model as follows. If the target sound for practice is German [e:], the anticipated errors of production will fall into three primary categories for American learners of German:

- the tendency to produce the vowel with too short duration;
- the tendency to diphthongize;
- the tendency to produce a vowel of the incorrect quality (either too high or too low).

Potential errors in the articulation [e:] and their acoustic manifestations are illustrated in the following chart.

CHART 1: Errors for [e:]

PRODUCTION ERRORS	ACOUSTIC MANIFESTATION
[e] too short pedagogically needs minimal [extended] length.	F_2 of < 1.5 sec.
[e^j]/[ε^j], etc. diphthongization	F_2 change of > ± 50 Hz (in > .15 sec)
quality errors:	
a. [$i^{(j)}$] too high (result of perception studies) with or without diphthongization	F_2-F_1 = >2K or F_2 = >2.5K (in <.15 sec)
b. [$\varepsilon^{(j)}$] too open/low with or without diphthongization	F_2-F_1 = <1.7K or F_2 = <1.9K (in >.15 sec)

4. PROCEDURE

For practical considerations, the utterance ['be: tən] is chosen for emulation. This choice facilitates computer analysis routines since the VOT of [b] corresponds closely to full consonantal release. Digitized samples of correct and incorrect production serve to develop the bases for the matching routines which perform the analysis functions.

Analysis at the first evidence of a harmonic wave (release spike/VOT of [b]) continues until the cessation of the harmonic wave (i.e., onset of [t]). The first and last 30 ms of the vowel are considered transitional and thus omitted from the LPC analysis.

The acoustic manifestations of the errors are processed in the sequence shown on the above chart.

- **Length.** If the wave length is less than 150 ms the computer does not process the signal and the student is prompted that the vowel is too short and is requested to produce the utterance again.

Only if the vowel is longer than 150 ms is the next routine enacted:

- **Diphthongization.** If a shift of more than 100 Hz is detected in the F_2 frequency during the steady state portion of the vowel, a message signaling diphthongization error is given and the student prompted to repeat the utterance.

If the vowel is produced with adequate length and without detectable diphthongization, then the last routine is enacted:

- **Quality.** If the figure F_2-F_1 is more than ±150 Hz. from a predetermined figure, an error message related to quality is given. Messages of too high or too low tongue position would be prompted depending on the type of subroutine triggered by the error.

Each time an error is made, the student is prompted as to the nature of that error according to the error routine triggered by the analysis of the digitized student response. Repetitions are elicited until no error matching routines are triggered at which time the utterance is deemed to have been correctly rendered.

5. IMPLEMENTATION

LPC analysis of a correct model and five potential error types were verified using the MacSpeech Lab II program indicated above. Based on these analyses, a set of criteria for the automated rating of these utterances was developed. Using a file exchange utility, the binary data was converted to the IBM PC format. A FORTRAN program was developed to 1) reverse the two data bytes necessary for software compatibility and 2) trim the data samples to isolate the start and end of the voiced portion of the utterance.

A prototype data analysis system was developed using MATLAB as a development environment. This system is compatible with virtually every popular platform. Data written as a 2-column ASCII array with FORTRAN is first imported into MATLAB using its "load" command. Then MATLAB scripts perform the analysis of the data and error detection according to a hierarchically arranged ranking order. Analysis is carried out of segments or windows of 30 ms in duration on the basis of estimated F_1 and F_2 frequencies.

Although MATLAB uses about 20 seconds to calculate a 256-point LPC, the calculations of formants and error criteria requires less than 5 seconds on a MAC II and a 12 MHz 285 PC-AT. By upgrading to a 386 environment with a 25 MHz clock speed, initiating onset of analysis at onset of voicing, and using "custom" software in assembly language to optimize performance, the error feedback time to students could be reduced to within the 1.5 second time window pedagogically needed.

6. CONCLUSION

Initial results have shown that the described model could be an effective pedagogical tool to enable error correction. Its proven functional success rests upon the predictability and normability of errors and specifically designed error-matching routines. This model does not depend upon full-spectrum matching routines. It may thus be the closest we can come to an automated phonetician at this time.

N.B. Much of the work for this paper was accomplished at IASCP at the University of Florida in Fall 1989.

7. REFERENCES

[1] FLANAGAN, J.L. (1972), *"Speech analysis, synthesis and perception"*, Berlin: Springer Verlag.

[2] REPP, B., et al. (1979), "Categories and context in the perception of isolated steady-state vowels", *Journal of experimental psychology, human perception and performance*, 5 (1) 129-145.

[3] ROACH, P., et al. (July 1990), "Phonetic analysis and the automatic segmentation and labelling of speech sounds," *Journal of the international phonetic association*, 20 (1), 15-21.

[4] WEISS, R. (1987), "Computer-assisted diagnosis of perceptual errors", *Proceedings of the XIth international congress of phonetic sciences*, Tallinn: Academy of Sciences of the Estonian S.S.R., 295-297.

[5] WEISS, R. and H-H. WAENGLER. (1985), *"German pronunciation: a phonetics manual"*, Bellingham: Western Washington University Press.

[6] ZWICKER, E., TERHARDT, E. AND E. PAULENS. (Feb. 1979), "Automatic speech recognition using psycho-acoustic models", *JASA*, 65 (2), 487-498.