A Visual Timing Tool for Language Training of Hearing Impaired Children

Uwe Meier
uwem@cs.cmu.edu

Jie Yang
yang+@cs.cmu.edu

Alex Waibel
ahw@cs.cmu.edu

Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

In this paper we present a timing tool for language training of hearing impaired children. Two major problems for these children to learn oral language are confusion of different phonemes and the timing within different words and syllables. In this paper we focus on developing techniques to overcome the timing problem. The techniques include timing error detection, visual feedback, and head tracking. The aim of the timing tool is to provide support to improve the speaking ability of the kids who are profoundly deaf and aided by hearing aids, cochlear implants, or a combination of both. We first describe the language training process in which the timing problem is addressed. We propose to use a modified dynamic time warping (DTW) algorithm to detect timing errors. The timing information is fed back to the user through a visual interface. Because the auditory stimulus is degraded, the visual feedback is extremely important to these kids. Two schemes can be used to produce the visual feedback. One scheme is to use the recorded video sequence of the kid to produce a corrected video replay; the other is to use a synthetic head to emphasize visual parts (lip-movement) that is wrong.

1. Introduction

Language training is a range of teaching and learning activities that help people acquire the abilities and skills to understand and produce spoken and written language. Language training can not only help the normal people learn to read, to learn a new language, but also help people with sensory, motor or learning disabilities to speak and understand their native language.

Visual speech provides complimentary cues to auditory speech. Thus, spoken language is more intelligible when the perceiver can view the talker's face. When the auditory stimulus is degraded (due to a hearing impairment or the acoustic environment), visual speech can improve intelligibility dramatically.

The research is currently supported by a NSF challenge grant. We are closely collaborated with the Oregon Graduate Institute and the University of California at Santa Cruz to develop interactive tools, technologies and applications for learning and language training with profoundly deaf children [1]. The goal of the project is to provide teachers, students, parents and other interested individuals with state of the art tools and technologies for interactive learning and language training. The systems runs under Windows 95 and Windows NT operation systems.

Figure 1 Hearing impaired children learning with the computer at the Tucker Maxon Oral School in Portland

Most speech recognition systems assign scores to hypotheses for acoustic acceptance/rejection of recognition. These scores, provide only a relative indication of goodness, and are limited to an acoustic evaluation of plausibility. Visual speech recognition will provide another dimension of valuation. Furthermore the tool must analyze the correctness of the timing during speaking and provide appropriate visual feedback to the students.

In this paper, we propose to use a modified dynamic time warping (DTW) algorithm [4] to detect timing errors. The timing information is fed back to the user through a visual interface. Two schemes have been implemented to produce the visual feedback. One scheme is to use the recorded video sequence of the kid to produce a corrected video replay; the other is to use a synthetic head to emphasize visual parts (lip-movement) that is wrong. A real-time face tracker automatically locates the user's face. We have developed a face and facial tracking library for Windows 95/NT and defined APIs for various applications based on the real-time tracking techniques developed previously [6].

2. Problem Description

Consider the following scenario for a language training process: a hearing impaired child sits in front of the computer and has to read some words out of a predefined list of words or repeat words spoken by a synthetic face (we use the Baldi

System [2]). During speaking the kid is recorded with a video camera that is connected to a PC. The face of the kid is automatically located and tracked by our face tracking software (See Figure 1). After recording, the speech is analyzed to detect the typical mistakes made by hearing-impaired people such as wrong timing during speaking. Then some visual feedback is produced to show where the mistakes where made.

2.1 Visual Input

Most current speech-reading systems require users to wear head-mounted cameras or reflective markers on their lips, or extract the relevant facial regions manually, this is especially working with children not practicable. We have developed techniques to achieve non-intrusive lip-reading and allow freedom of movement based on a real-time face tracker. Section 3.2 gives a short overview over the face tracking system we use to record the child's face.

2.2 Time Alignment

One of the main problems of the hearing impaired children is the timing within a word of different syllables or phonemes. Such time alignment can be more easily detected in acoustic speech signals. The idea is to use a modified dynamic time warping (DTW) algorithm. The basic DTW algorithm has to be modified the standard warping functions couldn't be used. This method is described in Chapter 3.1 in more detail.

We are currently working on a lip model to get more information from the visual signals. Once this work being finished, the model will provide additional feature to detected timing errors. Furthermore this model can be integrated in our lip-reading system to detect phoneme confusion errors.

2.3 Visual Feedback

A reason of the timing problem is that the hearing impaired children have nearly no feedback of their own speech. If the system could provide a visual feedback, it will be effectively reduce the timing error.

A natural way for such a feedback is to use the child's face to generate the visual feedback. After measuring the timing differences between test and reference samples, the system can generate a new image sequence with corrected timing using the child's face. The Implementation of this method is described in Section 3.3.

With the rewritten video the child can compare the time alignment in a natural way (his/her own face) but in these sequences there is still a lack of detailed information for some visual features (i.e. the correct lip position). Therefore we could use an artificial talking head (Baldi) to generate a correct outputs showing both correct and wrong timing sequences with more visual details. Baldi will be controlled by the results from timing tool and the lip-reading system [3]. Baldi can than produce output in both ways, the way that the child spoke the sentence and the way he/she should speak.

3. System Implementation

The following sections give a more detailed overview of some parts of the system, which are already implemented.

3.1 DTW Approach

For most speech recognizers is very difficult to identify timing errors during the recognition. However, the language training demands error detection. By investigating the problem of error detection, we found that the problem can be greatly simplified in the language training. In fact, we are not interested in recognizing what was spoken (we already know it in advance). More specifically, we are interested in which phoneme was spoken too fast or too slow in the timing problem. We then can solve the problem using the dynamic time warping (DTW [4]) algorithm.

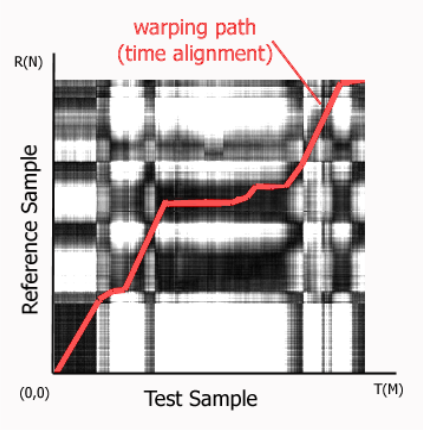

Figure 2: An example of using DTW algorithm for time alignment

The basic idea of the DTW algorithm is as follows. The goal is to find a warping path *w(t)* which maps the reference pattern *R(t)* onto the corresponding parts of the test pattern *T(t)*. The criterion for correspondence is that some measure of distance between the functions to be minimized by the mapping *w*. In this research, we use the Euclidean Distance between the MelScale coefficients of the acoustic samples as distance measure. The dynamic-time warping alignment problem is to find the optimal warping path *w(n)* which minimizes the accumulated distance D between test and reference patterns, subject to a set of path and endpoint constraints. Figure 2 shows the result of one of our test-samples. The x-aches represent the sample-points of the new-recorded test sample, the y-aches represent the pre-recorded reference sample. In the resulting *NxM* grid the Euclidean Difference of the MelScale Coefficients of the samples is shown (white=max difference, black=no difference). The warping path found by the algorithm is also showed in the same figure.

The constraints in the warping functions of the standard DTW are made to limit computing time (and errors) and are based on the observation, that 'normal' speaking people don't vary too much the speaking speed. But that's the reason why we can't

use this constraints but must modify the constraints to allow a much higher variety of changes in speaking speed.

3.2 Face and Facial Feature Tracking Library

It is well known that hearing impaired listeners and listening in adverse acoustic environments rely heavily on visual input to disambiguate among acoustically confusable speech elements.

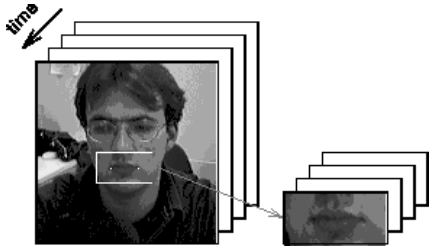

Figure 3. Basic idea of lip tracking

It has been demonstrated that visual information can enhance the accuracy of speech recognition. However, many current lip-reading systems require users to keep still or put special marks on their faces. We have developed a lip-reading system [5] based on a real-time face tracker [6]. The system first locates the face and then extracts the lip regions as shown in Figure 3.

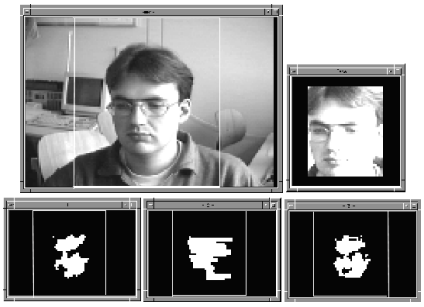

Figure 4. Face tracker results, upper row: camera image and found face, lower row: color classifier, motion classifier and combined color and motion classifier.

We have developed techniques of tracking human face and facial features in real-time [6]. Three types of models have been employed to track human faces. A skin-color model is used to register the face, a motion model is used to estimate image motion and to predict the location of the search window, as shown in figure 4. Finally a camera model predicts and compensates for camera motion (panning, tilting, and zooming). The system can track a person's face while the person moves freely (e.g., walks, jumps, sits down and stands up). The QuickTime movies of demo sequences in different situations and on different subjects can be found on our web site http://www.is.cs.cmu.edu/.

The real-time face and facial feature tracking systems were originally developed under UNIX platforms and were separated systems although they had some common components. Since the face and facial tracking systems have been serving as a base for many applications, a unified module system is desirable. In this research, we have developed a face and facial feature tracking library for both UNIX and Windows platforms based on the UNIX version of the systems. The library is arranged in such a way that it is very easy to configure an application system.

Several APIs (Application Programming Interface) have been defined to allow a user to quickly implement the face and facial tracking functions using C or the script language TCL/TK. The user can obtain coordinates of the face and/or facial features such as eyes, nostril, and lip-corners. Figure 5 shows a screenshot from a TCL/TK implementation. The following is the TCL/TK code is used to extract the lips region of a speaker.

```
## TCL/TK code for extracting a lip region
    package require tracker
    ## init Display
    set ft   [frame  .facetracker]
    set disp [canvas $ft.canvas -width 160\
            -height 120 -bg black]
    pack  $ft $disp -side top
    ## init Tracker
    image create photo tracker_photo
    tracker init tracker_photo
    $disp create image 0 0 -anchor nw \
                  -image tracker_photo
    ## main loop

    while {1} {
          ## process next image
          tracker next
          ## get results
          set facepos [tracker face]
          set lipspos [tracker lips]
          ## display results
          puts face position is $face
          puts lip position is $lips
}
```

3.3 Visual Feedback Generation

One way to produce the visual feedback is to rearrange the video sequence of the kid with correct timing. In the current implementation we rewrite the recorded video sequence based on the results of the DTW timing detection. While the child is speaking, the system records the face images based on the results from the face tracking system. The modified DTW algorithm is then used to detect the time alignment. If the child spoke too slow, some of the corresponding frames are skipped; and if the child spoke too fast, some frames are inserted by interpolation, so that the final sequence has the correct timing according to the reference sample. Figure 7 shows an example of the video rewriting. In the example, the speaker spoke frame too slow at the frame 3 (note that the frame number is not real number) and too fast at the frame 7. Accordingly, the original video sequence is rewritten, displaying a colored bar in every frame to visualize the results of the timing tool (green for correct, red for to slow, blue for to fast).

If the child speaks much too fast some smoothing must be performed to produce a more natural output. For the visual frames this will be done using the lip model. Since we have the exact lip position and model parameters of each frame and we know from the reference sample which phoneme was supposed

to speak and we can use some morphing technique to generate the missing frames.

Figure 5 Facial Feature Tracking Application for extracting lips-regions written with the API in TCL/TK

Using these two new generated video sequences, the child can look at its own face speaking in the correct way and in the way it actually spoke with visual additional information about the timing (the colored bars).

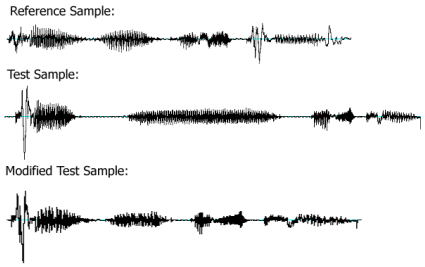

Figure 6 Modification of the Test Sample based on the output of the Timing Tool

Figure 6 shows the results of the algorithm on the acoustic waveform. The reference sample is the pre-recorded audio and the test sample the actual recorded. The modified test sample shows the results after modifying the waveform according to the results of our timing tool.

4. Conclusion

In this paper we have presented the ongoing research on a visual timing tool for language training of hearing impaired children. We described the problems and solutions for time alignment in the language training of hearing impaired children. The modified DTW algorithm has been used to detect timing errors. The timing errors are fed back to the child by a visual interface. We are currently improving the system and will soon test it in real language training process.

5. Acknowledgments

Support for this work has come from the NSF, under contract CDA-9726363, and the DARPA, under contracts N00014-93-1-0806 and N6601-97-C8553.

References

[1] R. Cole et al: *Intelligent animated agents for interactive language training*, Speech Technology in Language Learning, ESCA Workshop (STILL 98)

[2] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* Cambridge, MA: MIT Press (1998).

[3] U. Meier, W. Hürst, P. Duchnowski: *Adaptive Bimodal Sensor Fusion for Automatic Speechreading*, ICASSP 1996

[4] L.R. Rabiner S.E. Levinson: *Isolated and Connected Word Recognition – Theory and Selected Applications.* In: Reading in Speech Recognition, Edited by Alex Waibel & Kai Fu Lee, 1990, pages 115-153.

[5] R. Stiefelhagen, U. Meier, J. Yang, *Real-Time Lip-Tracking for Lipreading*, Eurospeech 97

[6] J. Yang, R. Stiefelhagen, U. Meier, A. Waibel: Visual *Tracking for Multimodal Human Computer Interaction*, CHI 98

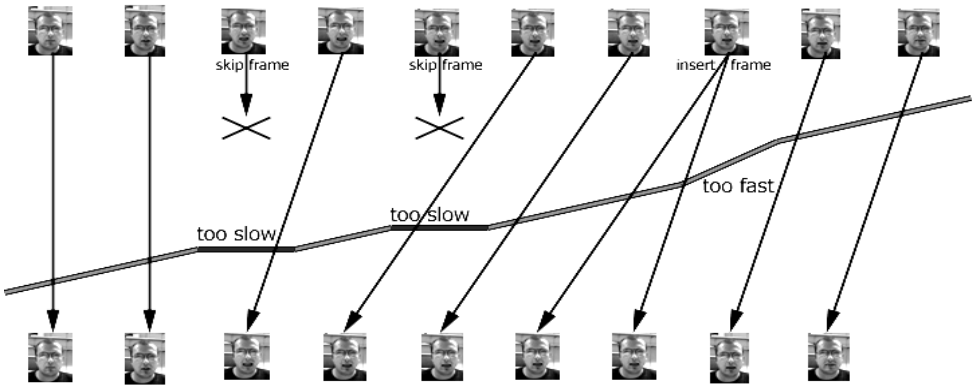

Figure 7 Example of rewriting the video sequences. The upper images are the original video frames, the lower images are the rewritten video frames. The line in the middle displays the results of the timing tool, in frame 3 the subject spoke too slow, in frame 7 too fast.