A Phoneme Probability Display for Individuals with Hearing Disabilities

Deb Roy and Alex Pentland
Perceptual Computing Group
MIT Media Laboratory
20 Ames Street, Rm. E15-388
Cambridge, MA 01239
(dkroy,sandy)@media.mit.edu

ABSTRACT

We are building an aid for individuals with hearing impairments which converts continuous speech into an animated visual display. A speech analysis system continuously estimates phoneme probabilities from the input acoustic stream. Phoneme symbols are displayed graphically with brightness in proportion to estimated phoneme probabilities. We use an automated layout algorithm to design the display to group acoustically confusable phonemes together in the graphical display.

Introduction

We are working on automatic speech analysis techniques which can be used to augment the communication abilities of hearing-impaired individuals. For example an individual might wear a portable device which includes a microphone to record the communication partner's speech, and some sort of tactile, visual or electrical stimulus output device which displays the speech after some preprocessing.

A basic design decision in building such a hearing aid is to determine how much interpretation of the speech signal will be left to the user versus the machine. At one end of the spectrum, the communication aid might extract only basic spectral characteristics of the input speech and present them to the user through tactile, visual or electrical stimulus (a cochlear implant is an example of such a device). Given a low level representation of the speech signal, the user is responsible for interpreting the spectral patterns as phoneme and ultimately word sequences. Such a device may complement information available to the user through other channels such as lip reading similar to cued speech [1].

On the other end of the spectrum, one might imagine a portable device which consists of a speech recognizer and a graphical display. To communicate with a hearing impaired user, the communication partner would talk into a microphone attached to the device. The speech recognizer would convert the speech to text which would be shown on a portable display.

Unfortunately state-of-the-art speech recognizers will produce high error rates when applied to open domain conversational speech. High accuracy recognition results which are typically reported for speech recognizers are achieved by limiting the domain (topic) of the speech. When the domain is known in advance, reliable vocabularies and statistical language models may be used to constrain the recognition task and vastly improve recognition accuracy. In the scenario presented above, if we wish to put no restrictions on the topic discussed using the communication aid, the speech recognizer may not rely on any domain dependent assumptions.

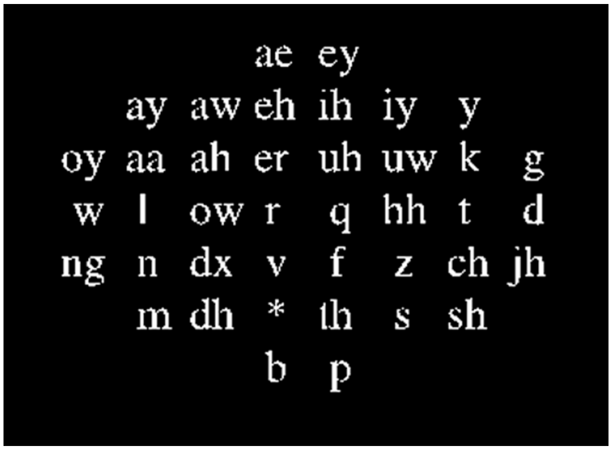
Our approach is to use automatic methods to estimate phoneme probabilities which are displayed graphically. The system makes no hard decisions about phoneme classification but instead displays all phoneme probabilities to the user allowing them to interpret sequences of probability estimates as word sequences. We use an automatic layout algorithm to design the graphical display with the goal of improving readability.

Phoneme Probability Estimation

Audio is sensed using a head mounted noise cancelling microphone (this would be worn by the person talking to the disabled user). The audio signal from the microphone is sampled at 16-bit 16 kHz analog and then processed using the Relative Spectral (RASTA) algorithm [3]. RASTA provides a spectral representation of the audio signal which is relatively invariant to background noise.

The RASTA coefficients are computed on 20ms frames of audio (recomputed every 10ms) and fed into a recurrent neural network (RNN) similar to the system described in [6] to produce phoneme probability estimates at a rate of 100Hz. The RNN has been trained using back propagation in time [10] on the TIMIT database [8] which is a database of phonetically labeled speech recorded from 630 adult male and female native English speakers from all the major dialect regions of the United States. The RNN produces a 40-dimensional phoneme probability vector (39 phoneme classes as defined by [5] and silence) every 10ms.

On the standard speaker independent TIMIT test set, the RNN recognizes phonemes with 68% accuracy. To act as a hearing aid, we need to display the output of the RNN in graphical



ae ey
ay aw eh ih iy y
oy aa ah er uh uw k g
w l ow r q hh t d
ng n dx v f z ch jh
m dh * th s sh
b p

Figure 1: Phoneme symbol layout which was generated automatically using simulated annealing

form. One option is to classify each frame of audio by selecting the phoneme with the highest probability estimate from the RNN and displaying the symbol of this phoneme (this is similar to the approach taken in [2] to display phonemes through a vibrotactile output display). Although this simplifies the information which is displayed to the user, it also introduces errors since the RNN misclassifies approximately three in ten phonemes. Instead we choose to display the probability of all 40 phonemes without forcing the system to make a hard classification. In the next section we describe the design of the display which attempts to display the probability vector in a natural way.

The Graphical Display

The probability vector produced by the RNN is connected to a graphical display. The display consists of a lattice of phoneme symbols as shown in Figure 1 (We use the same 39 phoneme classes as [4] and represent silence with an asterisk). Simulated annealing [7] was used to automate the phoneme layout so that phonemes with high confusability were placed in close physical proximity. Confusibility was measured by running the entire TIMIT test data set through the RNN and tabulating the classification errors made by the network for each phoneme. For example in Figure 1 the phonemes /t/ and /k/ are adjacent since they are often confused by the RNN. On the other hand /m/ and /g/ are placed on opposite sides of the display since they were rarely confused for each other.

To display a spoken utterance, the brightness of each symbol is set proportional to the probability estimate of that phoneme computed by the RNN. Figures 2 and 3 show a sequence of seven screen shots as the word “unjustified” is spoken. The first screen shows the display transitioning from the /ah/ to the /n/ phoneme in the first syllable of the word. Notice that the RNN is confident in its classification of /n/ (where there is significant confusion with only the phoneme /m/). In contrast there is less confidence in the identity of the first initial phoneme so the display lights up at least four symbols. The rest of the screen shots may be similarly interpreted to give the reader an idea of typical output from the display.

The resulting effect may be likened a spot light which sweeps over the display lattice. At instances when the output of the RNN is relatively confident (i.e. most of the probability mass of the vector is assigned to relatively few phoneme classes) the spot light is bright and focused on only one or a few phoneme symbols. When the network is less confident the spotlight becomes more diffuse and partially lights several phonemes. The goal is to not force a classification decision from the RNN but instead display all information from the RNN and leave the higher level interpretation of word sequences up to the user.

Summary and Future Directions

We have presented a system for converting continuous speech to an animated graphical display. Our emphasis was to provide some higher level analysis of speech (conversion to phoneme probabilities) without forcing the machine to make any hard decisions. This lets the user apply her knowledge of the situational context to constrain interpretation of the speech signal. We use a recurrent neural network (RNN) to estimate phoneme probabilities from continuous speech. To reduce the complexity of display a 40-dimensional phoneme probability vector, we designed a two-dimensional phoneme display which maps probabilities to brightness and arranges the phonemes so that confusable phonemes are placed close to one another. The display results in a “spotlight” effect which sweeps over the display and lights up phoneme groups over time. The size of footprint and intensity of the spotlight reflects the phoneme classification confidence of the underlying RNN.

In closing we mention some future directions for this work. Due to the early stage of this work we have not performed any usability tests although tests will be essential in assessing the practicality of our ideas. The system currently runs in real-time on an SGI R4400 workstation. Our ultimate goal is to implement the system on a wearable computer [9] thereby providing the hearing aid functionality in the user’s clothes and glasses. We also plan to to experiment with non-text symbols for the graphical display.

REFERENCES

1. O. Cornett. Cued speech. *American annals of the deaf*, 112:3–13, 1967.

2. E. M. Ellis and A. J. Robinson. Spiral: A vibrotactile based speech listening aid. In *Third International Conference on Tactile Aids, Hearing Aids and Cochlear Implants*, Miami, 1994.

3. H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, October 1994.

4. Kai-Fu Lee. *Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system.* PhD thesis, Computer Science Department, Carnegie Mellon University, 1988.

5. Kai-Fu Lee and Hsiao-Wuen Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, November 1989.

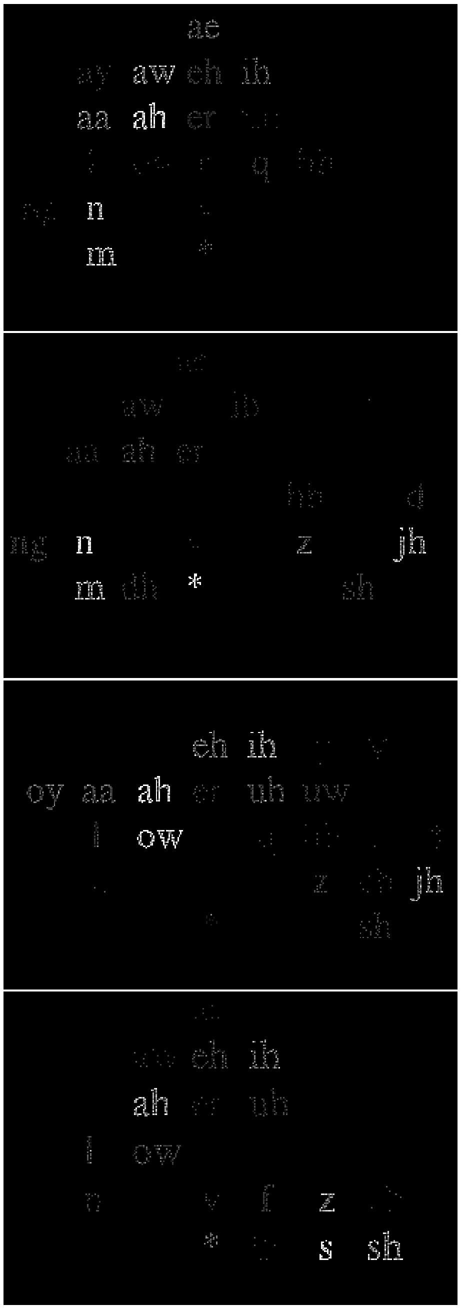


Figure 2: Screen shots of the phoneme display while showing the first two syllables of the spoken word *unjustified* .

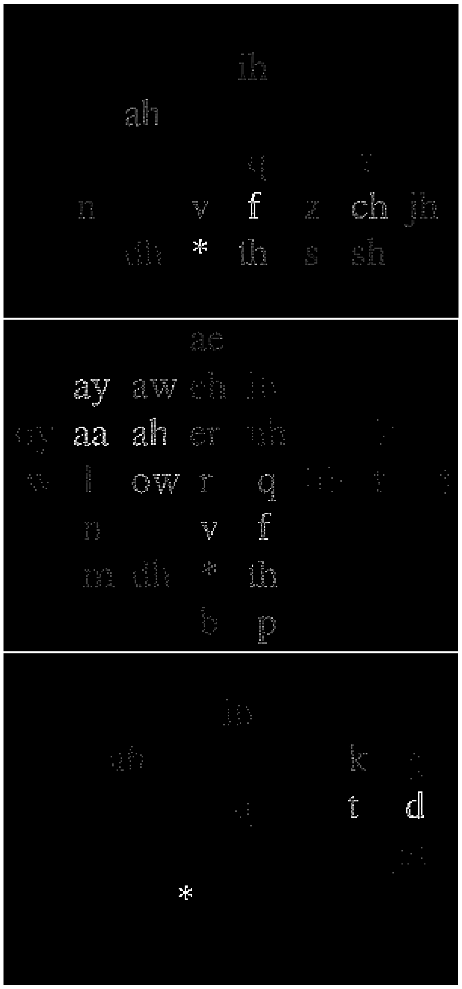


Figure 3: Screen shots of the final syllable in the word *unjustified* (notice that the speaker did not pronounce the *ti* portion of the word).

6. Tony Robinson. An application of recurrent nets to phone probability estimation. *IEEE Trans. Neural Networks*, 5(3), 1994.

7. C.D. Gelatt S. Kirkpatrick and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

8. Stephanie Seneff and Victor Zue. Transcription and alignment of the timit database. In *Proceedings of the Second Symposium on Advanced Man-Machine Interface through Spoken Language*, Oahu, Hawaii, November 1988.

9. Thad Starner, Steve Mann, Bradley Rhodes, Jeffrey Levine, Jennifer Healey, Dana Kirsch, Rosalind W. Picard, and Alex Pentland. Augmented reality through wearable computing. Technical report, MIT Media Lab Vision and Modeling Group, 1997.

10. Paul Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78:1150–1160, 1990.