Evaluating Automatic Speech Recognition Systems in Comparison with Human Perception Results Using Distinctive Feature Measures

Xiang Kong¹, Jeung-Yoon Choi², Stefanie Shattuck-Hufnagel²

¹Department of Computer Science, University of Illinois at Urbana-Champaign ²Speech Communication Group, Research Laboratory of Electronics, MIT

Abstract

- Using measures derived from linguistic distinctive features to evaluate automatic speech recognition (ASR) systems.
- Presenting error patterns in terms of manner, place and voicing, along with an examination of confusion matrices via a distinctive-feature-distance metric.

Motivation

Word error rate (WER) and phone error rate (PER) are the most popular methods. However, there are two drawbacks:

- The phone sequence for a produced word may vary widely, so that direct comparison of pronunciation variants becomes difficult.
- It is not possible to measure how similar two phones are to each other using phones as analysis units.

Experimental Setup

- Databases: TIMIT & LAFF VCV
- Conversion of Miller & Nicely results
- Human perception of LAFF VCV in noise: six syllables /aa-baa/, /aa-da-a/, /aa-ma-a/, /aa-ch-aa/ and /aa-sh-aa/, each of which was embedded in full-band white noise ranging from 30dB SNR to -20dB SNR in 10dB decrements.
- Hidden Markov Model (HMM)-based ASR system & Deep Neural Network (DNN)-based system detection of consonants and glides, selecting 3Q speech files from the LAFB database, each of which included a consonant or glide between two /aa/ phones, such as /aa-z-aa/, /aa-b-b/, etc. White noise at various levels from 40dB SNR to -20dB SNR at 10dB decrements were added to form the test data set.
- Evaluating system: All results were tabulated in terms of error types and into grey-scale confusion matrices.

Distinctive Feature Distance

The distinctive feature distance $D(i,j)$ between phonemes $i$ and $j$ can be written as $D(i,j) = \delta' = \delta(\delta(i,j)d(i,j))$, where $d(i,j)$ is the $f$th distinctive feature of the phoneme $i$, $f \in \{\text{vowel, glide, consonantal, }\ldots\}$, and $\delta(i,j)$ is the Kroenecker delta,

$$\delta(i,j) = \begin{cases} 1, & i = j \\ 0, & \text{otherwise} \end{cases}$$

In our study, the total number of distinctive features is 24.

Results

- Manner Place and Voicing error patterns
- Comparison of confusion matrices in white noise
- Distinctive-Feature-Distance results

Conclusion

- Human perception results show that place features are most susceptible to misperception in white noise, followed by manner features, then voicing features. The DNN-based system showed similar patterns, although with more errors.
- The HMM-based system had less consistency in the error patterns. In the confusion matrices, human perception results show that most errors occur near the diagonal regions, with blocking effects around -10dB SNR, which indicate correct voicing detections at that level.
- Point to the possibility of incorporating parameters specifically related to voicing, manner, and place into acoustic models, and/or incorporating distinctive-feature-distance measures as training criteria for clearer modeling of ASR systems to human perception patterns.

Future Work

- Extend this comparison to include results for larger databases and other types of noise, e.g. babble and bandpass noise.

Contact Information

- The authors want to thank Dr. Rachel Theodore and Nicholas Monto at the University of Connecticut-Storrs for their expertise in conducting the human listening experiments.

Graphs and Figures:

- Figure 1: Error patterns for human perception results
- Figure 2: Error patterns for Miller & Nicely results
- Figure 3: Error patterns for HMM-based system results
- Figure 4: Error patterns for DNN-based system results
- Figure 5: Confusion matrices for human perception results
- Figure 6: Confusion matrices for Miller & Nicely results
- Figure 7: Confusion matrices for HMM-based system results
- Figure 8: Confusion matrices for DNN-based system results
- Figure 9: Plot of distinctive-feature-distance results of human perception and HMM- and DNN-based ASR systems