EVALUATING AUTOMATIC SPEECH RECOGNITION SYSTEMS IN COMPARISON WITH HUMAN PERCEPTION RESULTS USING DISTINCTIVE FEATURE MEASURES

Xiang Kong[1], *Jeung-Yoon Choi*[2], *Stefanie Shattuck-Hufnagel*[2]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign
[2]Speech Communication Group, Research Laboratory of Electronics, MIT
xkong12@illinois.edu, jyechoi@mit.edu, sshuf@mit.edu

ABSTRACT

This paper describes methods for evaluating automatic speech recognition (ASR) systems in comparison with human perception results, using measures derived from linguistic distinctive features. Error patterns in terms of manner, place and voicing are presented, along with an examination of confusion matrices via a distinctive-feature-distance metric. These evaluation methods contrast with conventional performance criteria that focus on the phone or word level, and are intended to provide a more detailed profile of ASR system performance, as well as a means for direct comparison with human perception results at the sub-phonemic level.

***Index Terms*—** performance evaluation, error patterns, confusion matrices, distinctive-feature-distance metric

1. INTRODUCTION

Conventional automatic speech recognition (ASR) systems model and evaluate the speech recognition process as follows. Acoustic measurements can be expressed as a sequence $X = \{x_1, x_2, ..., x_t, ..., x_T\}$, and the true word sequence as $W = \{w_1, w_2, , w_n, , w_N\}$, so that the optimum estimated sequence is $\widehat{W} = \arg\max P(W|X)$, or more generally, with a scoring function, $S(W, X) => \widehat{W} = \arg\max P(W|X)$. In order to compare the difference between W and $\widehat{W}$, the most commonly used evaluation measure is the word error rate (WER) [1]. The general form of an error rate is given by error rate = (deletions + insertions + substitutions) / (entries), and accuracy is (1 error rate). For different recognition tasks, other evaluation measures can be applied. For binary classification applications, such as keyword spotting, the F-score may be used, which is calculated as the harmonic mean of the precision and recall. To evaluate the word accuracy of translation from one language to another, the BLEU (Bilingual Evaluation Understudy) method uses the correspondence of the translation between the machine and humans, and requires word boundaries. Similarly, the ROUGE method evaluates word accuracy by comparing the translation from a machine to a human-produced summary [2] [3].

A common characteristic of the evaluation methods mentioned above is that the metrics are related to word-level matches. A drawback of evaluation measures at the word level is that they capture overall system performance, combining the results of acoustic and language models. A simple way to isolate the performance at the acoustic level is to evaluate at the phonemic level, using a unit such as the phone, leading to phone error rates (PER), which are generally lower than the WER. However, using PER to describe performance at the acoustic level is still problematic in at least two aspects. First, the phone sequence for a produced word may vary widely, especially for common function words and for word sequences susceptible to reduction, so that direct comparison of pronunciation variants becomes difficult. Second, it is not possible to measure how similar two phones are to each other using phones as analysis units. The first problem may be solved by restricting the test sequences to utterances with no pronunciation variants; an example would be VCV (vowel-consonant-vowel) syllables. The second issue can be addressed by describing phoneme level units at a sub-phonemic level, by using linguistic distinctive features [4]. For example, the phonemes **/f/** and **/s/** are both **[+consonantal]**, **[-sonorant]**, **[+continuant]**, but **/f/** is **[+labial]**, while **/s/** is **[+alveolar]**.

In line with such research, this paper describes methods for evaluating system performance using distinctive feature measures, for a detailed analysis of results at the sub-phonemic level. This approach enables descriptions of the non-uniform effects on recognition results from perturbations of the input speech, such as ambient noise, or from non-standard speaker characteristics, such as foreign accents. This approach also allows direct comparison with human perception results, and may provide directions for modeling human perception patterns more closely in ASR systems. Accordingly, in this paper, we observe the acoustic-level performance of two types of ASR systems perturbed with additive white noise in comparison with human perception results, using VCV syllables. ASR systems based on Hidden Markov Models (HMMs) and Deep Neural Networks (DNNs) are examined. In contrast to conventional evaluation meth-

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

ods, performance is analyzed by classifying error patterns related to manner, place and voicing, and by examination of confusion matrices via a distinctive-feature-distance metric.

2. METHODS

2.1. Databases

The TIMIT database was used for training the ASR systems, and the LAFF VCV database with additive noise was used as the test set. TIMIT [5] contains broadband recordings of 630 speakers of 8 major dialects of American English, each reading 10 phonetically rich sentences, and includes time-aligned orthographic, phonetic and word transcriptions, as well as a 16kHz speech waveform file for each utterance. LAFF VCV [6] is a collection of vowel-consonant-vowel syllables recorded at the MIT Speech Communication Group from 2 male speakers and 1 female speaker who were native speakers of North American English. The syllables were formed from 6 vowels and 26 consonants, resulting in utterances such as **/aa-b-aa/**, etc. Full-band white noise was added to the VCV utterances, to produce test stimuli with various signal-to-noise ratios.

2.2. Conversion of Miller & Nicely results

First, the results from the confusion matrices in Miller and Nicelys paper [7] on human perceptual confusions in noise were converted to manner, place and voicing errors as well as grey-scale confusion matrices, for more direct comparison with other results. Manner errors involve errors between vowels, glides = {w, y, r, l, h}, nasals = {m, n, ng}, fricatives = {f, th, s, sh, v, dh, z, zh}, stops = {p, t, k, b, d, g}, and affricates = {ch, dj}. Place errors indicate errors between labials = {m, f, v, p, b}, dentals = {th, dh}, alveolars = {n, s, z, t, d}, palatals = {sh, zh, ch, dj}, and velars = {ng, k, g}. Voicing errors are between unvoiced = {f, th, s, sh, p, t, k} and voiced = {all others}. In the grey-scale confusion matrices shown in Figs. 5 - 8, the magnitude of the entry in each cell corresponds to the darkness of the cell (e.g. 10 out of possible 10 trials would be 100%, or black). The additive white noise levels range from 12dB SNR to -18dB SNR, in 6 dB SNR decrements. In the following results, the entries in the confusion matrices are in the order of **/p, t, k, f, th, s, sh, ch, b, d, g, v, dh, z, zh, dj, m, n, ng, w, y, r, l, h/**, and No Response, with entries skipped if not present. Unvoiced sounds appear before voiced sounds, with stops preceding fricatives, and nasals and glides are at the end. This placement facilitates the visual analysis of error patterns in the grey-scale confusion matrices.

2.3. Human perception of LAFF VCV in noise

Next, human listening experiments were carried out with selected LAFF VCV data. The stimuli consisted of the following six syllables: **/aa-b-aa/**, **/aa-d-aa/**, **/aa-s-aa/**, **/aa-m-aa/**, **/aa-ch-aa/**, and **/aa-sh-aa/**. Each syllable was embedded in full-band white noise ranging from 30dB SNR to -20dB SNR in 10dB decrements. Twenty adults (17 females, 3 males) between 18 and 31 years of age (mean = 22, SD = 3) were recruited for participation in the experiment. All participants were monolingual speakers of American English and had no history of speech, language, hearing, or neurological disorders according to self-report. Each participant heard 360 syllables (6 consonants × 10 repetitions × 6 SNRs). The syllables were blocked by SNR, with block order randomly determined for each participant, and syllable order randomized within each block. Participants were directed to provide a response even when they were not completely sure of the item. Stimulus presentation and data collection was controlled using the SuperLab software [8].

2.4. Hidden Markov Model(HMM)-based ASR system

As an example of a widely-used HMM-based ASR system, we selected the standard HTK system [9]. Standard training methods for the HTK system were applied to obtain single-phone models (6 states with non-emitting first and last states/8-mixture Gaussian mixture models), and HInit and HRest procedures were used to initialize models before running HERest for model optimization [10]. The results showed 67% correct phone recognition for the entire TIMIT test set, and 65% for the TIMIT core test set. Error rates reported in the literature similarly ranges from 46% to 55%, in line with our error rates. For our study, we focused on detection of consonants and glides, selecting 36 speech files from the LAFF database, each of which included a consonant or glide between two "**aa**" phones, such as **/aa-z-aa/**, **/aa-b-aa/**, etc. White noise at various levels from 40dB SNR to -20dB SNR at 10dB decrements were added to form the test data set [11]. (A subset of these files was selected for use in the human perception tests in Section 2.5.) Acoustic feature files were obtained using Hcopy, the consonants were detected using HVite. The HTK results were matched to the true sequence (minimum edit distance match), and the results were tabulated in terms of error types and into grey-scale confusion matrices.

2.5. Deep Neural Network(DNN)-based ASR system

For the DNN training [12], 24 Mel-scale log filter banks are extracted as input features. There are five hidden layers, each of which has 2048 nodes, and additionally there is a softmax output layer; we use backpropagation to tune the weights. Overall, the DNN system was implemented using the KALDI toolbox [13], and trained on the TIMIT TRAIN set, resulting in a PER of 30.87% on the TIMIT TEST set.

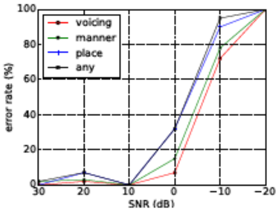


Fig. 1. Error patterns for human perception results

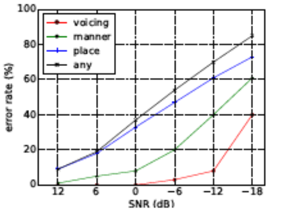


Fig. 2. Error patterns for Miller & Nicely results

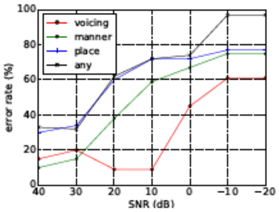


Fig. 3. Error patterns for HMM-based system results

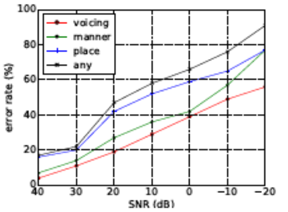
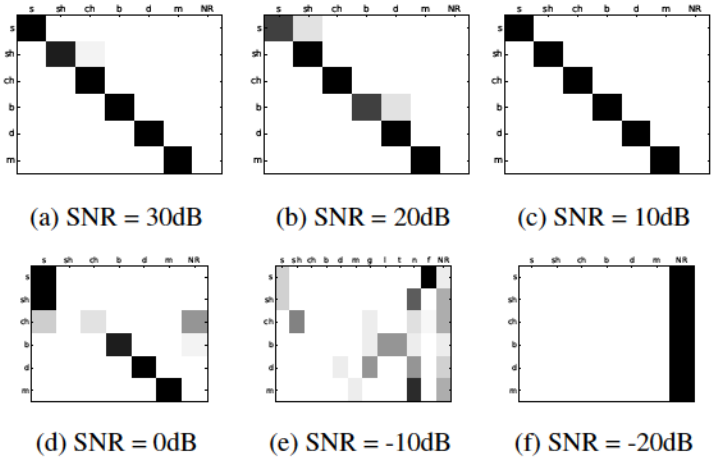


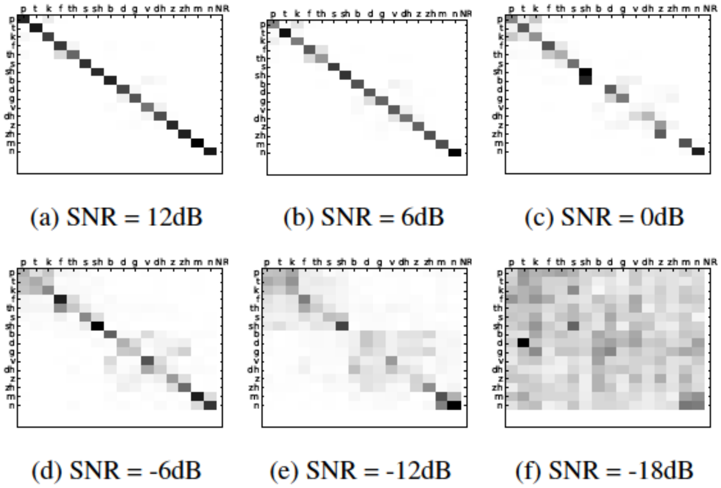
Fig. 4. Error patterns for DNN-based system results



(a) SNR = 30dB (b) SNR = 20dB (c) SNR = 10dB

(d) SNR = 0dB (e) SNR = -10dB (f) SNR = -20dB

Fig. 5. Confusion matrices for human perception results



(a) SNR = 12dB (b) SNR = 6dB (c) SNR = 0dB

(d) SNR = -6dB (e) SNR = -12dB (f) SNR = -18dB

Fig. 6. Confusion matrices for Miller & Nicely results

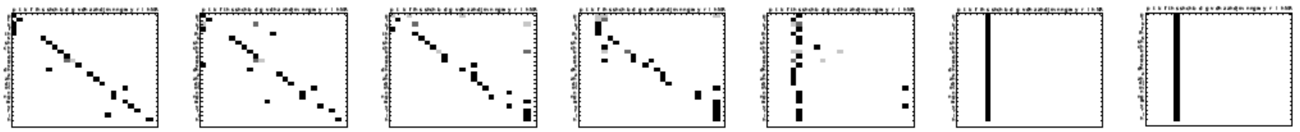
3. RESULTS

3.1. Manner Place and Voicing error patterns

Human perception results for LAFF VCV database files are shown in Fig. 1. The results show that as SNR decreases error rates increase, particularly after 0 dB SNR. Also, place errors are most prevalent, followed by errors in manner, then voicing. Next, the results converted from the Miller and Nicely paper are shown in Fig. 2. Similar to the results for the LAFF VCV database, errors increase monotonically, with place errors highest, followed by manner errors, then voicing errors. In comparison, results for HMM phone detection in the LAFF VCV database are shown in Fig. 3. In contrast with the results for human perception, it is not always the case that error rates increase monotonically as SNR decreases. Also, the relative degree of errors in place, manner and voicing are not consistent as in the human perception results. Nevertheless, it can be seen that place and manner errors are mostly higher than voicing errors, although the total number of errors are higher than for human listeners. The results for the DNN in Fig. 4 show error patterns that are more similar to human perception results, with the most place errors and the least voicing errors. However, the overall error rates are higher and increase starting at higher SNRs.

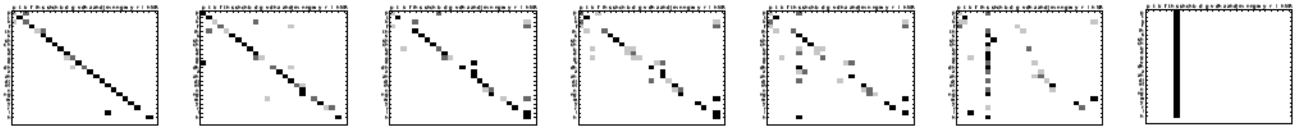
3.2. Comparison of confusion matrices in white noise

In order to examine error patterns in more detail, grey-scale confusion matrices for the detection results were constructed. The human perception results from the LAFF VCV database are shown in Fig. 5. As shown in the confusion matrices, in clean speech, most responses are on the diagonal, i.e. correct. In the next-to-last figure, the responses are spread out from the diagonal, indicating a more random detection pattern. In the final figure, all responses are given as ”No Response”, i.e. the human listeners were unable to answer. A similar pattern is seen for the Miller & Nicely responses in Fig. 6, in that lower noise levels result in more diagonal entries, while higher noise levels result in a more random error pattern. However, the next-to-last confusion matrix shows that even at a low SNR of **-12dB**, the errors appear in blocks, which indicate common voicing features. For the confusion matrices for the HMM results shown in Fig. 7, the overall error rates are higher than for the human perception results, even at the higher SNR levels. Moreover, the results do not exhibit the characteristic blocking patterns that indicate similar voicing detections at noise levels around **-10dB** SNR. And finally, it can be seen that around **0dB** SNR or lower, most re-



(a) SNR = 40dB (b) SNR = 30dB (c) SNR = 20dB (d) SNR = 10dB (e) SNR = 0dB (f) SNR = -10dB (g) SNR = -20dB

Fig. 7. Confusion matrices for HMM-based system results



(a) SNR = 40dB (b) SNR = 30dB (c) SNR = 20dB (d) SNR = 10dB (e) SNR = 0dB (f) SNR = -10dB (g) SNR = -20dB

Fig. 8. Confusion matrices for DNN-based system results

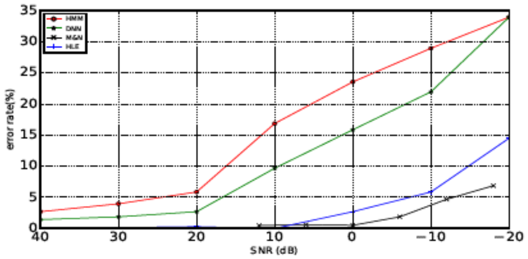


Fig. 9. Plot of distinctive-feature-distance results of human perception and HMM- and DNN- based ASR systems

sponses are given as **/s/**, which diverges from the human perception results. For the DNN results in Fig. 8, we can see that results are more similar to human perception results. However, we do not observe the characteristic blocking effect, and the responses at the lowest SNRs include a majority of **/s/** responses, similar to the HMM results. The distinctive feature distance $D(i, j)$ between phonemes i and j can be written as $D(i, j) = \sum_f (1 - \delta(d(i, j)d(j, f)))$, where $d(i, f)$ is the fth distinctive feature of the phoneme i, $f \in$\{vowel, glide, consonantal, ...\}, and $\delta(i, j)$ is the Kronecker delta, $\delta(i, j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$. In our study, the total number of distinctive features is 24. The normalized distinctive-feature-distances of the confusion matrices above are plotted in Fig. 9, which represents the sub-phonemic recognition performance of the ASR systems in comparison with the human perception results.

4. CONCLUSIONS

This study presents an evaluation method for sub-phonemic analysis of the effects of additive noise on two types of ASR systems, in direct comparison with human perception results. Results were compared in terms of manner, place and voicing error patterns, as grey-scale confusion matrices, and as distinctive-feature-distances. 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. In contrast, 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. Meanwhile, in the DNN- and the HMM-based systems, errors converge to the sound **/s/** at higher noise levels, more randomized errors occur at higher SNR levels, and the blocking effect from voicing detection is not observed. The distinctive-feature-distances of the confusion matrices summarize the discrepancy of the ASR system performance from human perception results at the sub-phonemic level.

These results 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 closer modeling of ASR systems to human perception patterns. Further work aims to extend this comparison to include results for larger databases and other types of noise, e.g. babble and bandpass noise, which can be expected to produce perturbation patterns that are different from that from white noise for human perception and ASR systems.

5. ACKNOWLEDGEMENTS

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.

6. REFERENCES

[1] Xiaodong He, Li Deng, and Alex Acero, “Why word error rate is not a good metric for speech recognizer training for the speech translation task?,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5632–5635.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[3] Chin-Yew Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out: Proceedings of the ACL-04 workshop*. Barcelona, Spain, 2004, vol. 8.

[4] Morris Halle, “Phonological features,” *International encyclopedia of linguistics*, vol. 3, pp. 207–212, 1992.

[5] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, “Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1,” *NASA STI/Recon technical report n*, vol. 93, 1993.

[6] MIT Speech Communication Group, “LAFF VCV Syllables,” `http://hdl.handle.net/1721.1/25142`, 2005, [Online; accessed 19-Aug-2005].

[7] George A Miller and Patricia E Nicely, “An analysis of perceptual confusions among some english consonants,” *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[8] Cedrus Corporation, “ SuperLab 5,” `http://cedrus.com/superlab/manual/`, 2014.

[9] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al., “The htk book,” *Cambridge university engineering department*, vol. 3, pp. 175, 2002.

[10] Philip C Woodland and Steve J Young, “The htk tied-state continuous speech recogniser.,” in *Eurospeech*, 1993.

[11] JE Hawkins Jr and SS Stevens, “The masking of pure tones and of speech by white noise,” *The Journal of the Acoustical Society of America*, vol. 22, no. 1, pp. 6–13, 1950.

[12] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[13] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., “The kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.