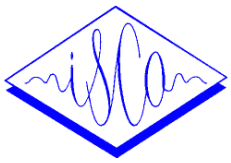

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

UTTERANCE VERIFICATION IMPROVES CLOSED-SET RECOGNITION AND OUT-OF-VOCABULARY REJECTION

*Don Colton** *Mark Fanty* *Ron Cole*

email: Don_Colton@cse.ogi.edu http://www.cse.ogi.edu/CSLU/
Center for Spoken Language Understanding, Oregon Graduate Institute
P. O. Box 91000, Portland, Oregon 97291-1000 USA

ABSTRACT

We report on utterance verification of putative recognitions in both open-set and closed-set recognition tasks using telephone speech. For open-set recognition, we report on rejection of out-of-vocabulary utterances. In a two-keyword task ("male" and "female") using 50% out-of-vocabulary utterances, utterance verification reduced errors by 60%, from 12% to 4.8% compared to our baseline rejection strategy. For closed-set recognition, we report on re-ordering the N-best hypotheses. In a 58-phrase task, utterance verification reduced closed-set recognition errors by 30%, from 6.5% to 4.5%.

1. INTRODUCTION

Recognition based on the combination of phonetic likelihoods from short fixed-width frames is the dominate paradigm for speech recognition systems. While this approach has numerous advantages, it is reasonable to think that better word-level recognition is possible using whole-word classifiers. Building such recognizers presents a number of difficulties, such as finding word boundaries before performing the classification, and collecting enough data to train the classifiers.

In this paper we report results on experiments with a two-pass strategy. The first pass uses a frame-based recognizer. The output is the recognized word (putative hit) or a list of the top N recognized words, along with the phonetic segmentation derived from backtrace information. This effectively solves the segmentation problem. For our experiments, ample training data was available for the entire vocabulary.

*This research was supported in part by a National Science Foundation Graduate Fellowship, and grants from U S WEST, the Office of Naval Research, NSF and ARPA.

Given a putative match between a test utterance and a reference phrase, we verify (i.e. confirm or deny) this match using word-specific classifiers. These are neural networks with input features describing the whole word. Combining reclassificatiion with an N-best recognizer allows us to improve recognition accuracy if the utterance verification score is more reliable than the initial recognition score. We can also reject out-of-vocabulary utterances by rejecting the entire set of top-scoring matches from the N-best list.

This paper extends prior work at the Center for Spoken Language Understanding (CSLU) on two-pass Alphabet recognition [1]. In the alphabet system, the frame-based first pass provides letter and broad-phonetic boundaries. The second pass uses an extensive set of knowledge-based features specifically designed for the alphabet. The second-pass classifier has 27 outputs: the 26 letters and an output for "not a letter" which was trained on false positives from the first pass in a development set (mostly noise, not extraneous speech). The second pass yielded much better recognition than we could achieve with a frame-based recognizer alone. The work presented here differs in several ways: the classifiers are word specific, so there are two outputs: word and not-word. This contrasts with having the whole vocabulary in a single network. Also, the feature set is generic and not based on careful study of the vocabulary.

Our work also extends that of Mathan and Miclet [2]. They used word-specific neural networks to reclassify putative hits in an isolated word recognizer. Their feature vector included duration, average energy and the average first Mel frequency coefficient for each segment in the trace of the first-pass recognition as input features. We extend this work by examining a variety of feature bundles, and by combining reclas-

sification with an N-best search list to improve keyword recognition accuracy.

In all our experiments, telephone speech was used. The speech was digitally sampled at 8000 Hz. For all our corpora, calls are serially numbered as they arrive, and are apportioned into training (60%), development test (20%), and final test (20%) sets according to the last digit of the serial number.

2. THE FRAME-BASED CLASSIFIER

For both experiments, the first pass is a frame-based classifier which uses a neural network to estimate phoneme probabilities. Speech analysis is seventh order Perceptual Linear Prediction (PLP) analysis [3], which yields eight coefficients per frame including energy. The analysis window is 10 msec and the frame increment is 6 msec. The inputs to the neural network are 56 PLP coefficients from a 160msec window around the frame to be classified. The outputs of the network correspond to the phonetic units of the task. For the male/female task the net has only six outputs. For the 58-word task, we used a context-dependent net with sub-phoneme units [4] and there were several hundred outputs.

The best alignment of a vocabulary word with the neural network probability estimates is found using a Viterbi search. Background sounds are modeled with a simple garbage model [5] which increases robustness and provides some wordspotting ability. This makes out-of-vocabulary rejection more difficult, as the vocabulary word need only align with part of the extraneous speech.

3. OUT-OF-VOCABULARY REJECTION

Our first experiment sought to identify and reject out-of-vocabulary utterances using a second-pass, whole-word classifier. The task was gender recognition which consisted of two words: "male" and "female." This is an easy task for which the frame based classifier does very well, but it is fairly difficult for rejection because the target words are so short.

All speech data in this experiment are from the OGI Census corpus [6]. We used gender utterances and last name utterances. The gender utterances consist of more than 2000 responses to the prompt "What is your sex, male or female?" Of these, roughly 70% were the word "female" (including a few examples spoken by males!) and 30% were the word "male." The last name utterances consist of responses to the prompt "Please say your last name."

3.1. Baseline System

The baseline system was a frame-based neural network recognizer for the two words "male" and "female." This recognizer was developed for and used in the OGI Census system [7]. When in-vocabulary utterances are used, the baseline system's accuracy is 99.5%. To detect low-confidence recognitions, the baseline system takes the ratio of the top two recognizer scores, and compares this to an optimized threshold.

3.2. Second Pass Rejection

Our approach is to take the Viterbi backtrace to identify the start and end times for each phoneme of the putative utterance. We then collect features based on this time alignment, and use them to train two new networks (one each for "male" and "female"). The new networks produce two outputs: "confirm" and "deny."

The training set contained as many negative examples as positive. The Census corpus contained very few extraneous utterances, so we ran the male-female recognizer on the Census corpus of last names (family surnames), forcing each to be recognized as "male" or "female," and used these as negative inputs for training and testing.

The "female" utterance verifier was trained using 2000 examples, and (due to less available data) the "male" utterance verifier was trained using 1400 examples. In each case half of the training examples represented correct putative hits (drawn from the gender corpus) and half represented incorrect putative hits (drawn from the last name corpus). Similarly, half of the test set was "male" or "female" and half was last names. Using the Viterbi backtrace from the first-pass recognition, we identified word and phoneme boundaries (three phonemes for "male" and five for "female"). We then looked at the following feature combinations:

1. **[du]** Phoneme durations alone.

2. **[en]** Phoneme center-frame energy alone.

3. **[du.en.+]** Phoneme durations, phoneme center-frame energies, plus the energy in the frame 50 msec before and the frame 50 msec after the word.

4. **[du.10p]** Phoneme durations plus PLP from ten frames located at 5%, 15%, 25%, ..., and 95% across the word.

5. **[du.5p]** Phoneme durations plus PLP from five frames located at 5%, 25%, 45%, 65%, and 85% across the word.

6. **[du.sp.+]** Phoneme durations, PLP from the center-frame of each phoneme, plus the PLP from the frame 50 msec before and the frame 50 msec after the word.

3.3. Results

Setting the rejection threshold for the best overall performance on a development set which had an equal number of examples of in-vocabulary and out-of-vocabulary speech, the best we could do with the baseline system was 88% overall.

All but one of the feature sets used for second pass classification scored better. Phoneme durations alone [du], a very small number of input features, do quite well. Durations and energies [du.en.+] scored about the same as durations alone. Energies alone [en] scored much worse. As expected, durations plus PLP from the center of each phoneme [du.sp.+] scored best. Sampling PLP equally across the word [du.10p] [du.5p] did not work as well as using the phonetic boundaries from the first pass.

The following table shows the utterance verification accuracy for each of the six feature vector sets, for each of the two keywords. An overall (weighted) average is also shown, and this is compared to the baseline accuracy of 88% to give a measure of error reduction.

Results	overall	gain
1. du	.928	.400
2. en	.803	(.642)
3. du.en.+	.927	.392
4. du.10p	.941	.508
5. du.5p	.926	.383
6. du.sp.+	.952	.600

In each case, putative hits for "female" were reclassified more accurately than those for "male." This may be due to the smaller training set for "male" or because there are fewer phonemes on which to base a decision.

4. IMPROVED CLOSED-SET RECOGNITION

In our second experiment, we used reclassification to re-order an N-best hypothesis list in order to improve recognition accuracy. The closed set consisted of 58 words and phrases in the telephone services domain. Phrases varied in length from two to twenty-three phonemes. Our task was to reclassify the top three choices and possibly change the identity of the recognized utterance.

More than 1000 callers said each of the 58 target words or phrases. Each utterance was verified by a human listener, and mistakes (for example, the wrong phrase or a partial phrase) were deleted from the corpus. There was no extraneous speech.

4.1. Baseline System

The baseline system was a frame based neural network classifier plus Viterbi search. Left and right context dependent modeling, with categories chosen specifically for this vocabulary, resulted in over 500 outputs. Each base phoneme was divided into three parts: left-context dependent, center, and right-context dependent. Using only in-vocabulary test utterances, with each of the 58 phrases equally likely, the accuracy is 93.5%. When there is an error, the correct phrase is often near the top of the N-best list. This is what prompted us to try a second pass classifier.

4.2. Second Pass Rescoring

We trained a neural network for each of the 58 keywords using a subset of the data. An equal number of positive and negative examples were used for each. Negative examples were chosen from the utterances for which the target word appeared high in the N-best list (i.e. we selected the more easily confused utterances from within the 58-word vocabulary).

Using our experience from the first experiment, we based our feature vector on the segmentation from the Viterbi backtrace on each putative hit in the N-best list. The following features were used for utterance verification:

- The average per-frame Viterbi score for the entire word (from the first pass recognizer).
- The average per-frame Viterbi score for each sub-phonetic segment.
- The duration of each sub-phonetic segment.
- The PLP from the center of the middle (context-independent) segments.
- The PLP from the frame 50 msec before and the frame 50 msec after the word.

By reviewing the development test scores, a manually optimized threshold was developed to select the best match from the reclassification scores of the top three outputs of the N-best classification: If scores one and two were both below 0.1, and score three was above 0.5, then the third match was selected (this was rare). Otherwise, if score two was 1.7 times greater than score one, the second match was selected. Otherwise the first match was selected.

4.3. Results

On the final test set, the error rate without utterance verification was 6.5%. The verification step error rate was 4.5%, which is a 30% improvement. It is interesting to note that when an early version of the first-pass recognizer was below 90% accuracy, the verification improved the performance to about 95%. As the first pass improved, the net result after the verification held steady.

5. CONCLUSIONS

Word-based reclassification showed promise in both experiments. For rejection, it worked better than our default scheme of using ratios. Although the default was no doubt not the best possible one-pass rejection strategy, the second pass could probably be improved as well. For example, (in the first experiment) no features based on the phonetic probabilities from the first pass were used. The biggest drawback of this approach is the large amount of training data needed to build the classifiers. It is possible to formulate word acceptance as a vocabulary-independent classification problem based on features sets which can be defined for any word. This will be investigated in the future. Vocabulary-specific reclassification will be reserved for special applications such as number recognition.

References

[1] Mark Fanty, Ronald A. Cole, and Krist Roginski. English alphabet recognition with telephone speech. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 4*, Denver, 1992. Morgan Kaufmann.

[2] Luc Mathan and Laurent Miclet. Rejection of extraneous input in speech recognition applications, using multi-layer perceptrons and the trace of HMMs. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-91)*, pages 93–96, Toronto, May 1991.

[3] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[4] Etienne Barnard, Ronald A. Cole, Mark Fanty, and Pieter Vermeulen. Real-world speech recognition with neural networks. In *Applications and Science of artificial neural networks*, volume 2492, pages 524–537. SPIE, April 1995.

[5] Jean-Marc Boite, Hervé Bourlard, Bart D'hoore, and Marc Haesen. New approach towards keyword spotting. In *Proceedings of the Third European Conference on Speech Communication and Technology (EUROSPEECH-93)*, pages 1273–1276, Berlin, September 1993.

[6] Ronald Cole, Mark Fanty, Mike Noel, and Terri Lander. Telephone speech corpus development at CSLU. In *Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP-94)*, Yokohama, Japan, September 1994.

[7] Ronald Cole, David G. Novick, Mark Fanty, Pieter Vermeulen, Stephen Sutton, Dan Burnett, and Johan Schalkwyk. A prototype voice-response questionnaire for the U.S. census. In *Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP-94)*, Yokohama, Japan, September 1994.