Relating Frame Accuracy with Word Error in Hybrid ANN-HMM ASR

*Michael L. Shire**

International Computer Science Institute
Berkeley, California USA
shire@icsi.berkeley.edu

Abstract

Frame accuracy is a common and natural summary statistic to use in neural-network-based ASR. It is often used as an indication of the performance of the neural network probability estimator and in the stopping criterion during its training. Though considered an important factor for word recognition, the frame accuracy presents an incomplete and sometimes deficient indicator of performance for the overall task of word recognition, as with many such summary statistics. Many in the ASR community have seen instances where an improvement in the acoustic posterior probability estimation yielded a disappointing effect on word recognition. We conducted experiments in an effort to illustrate some of the variability in word-recognition performance associated with frame accuracy. Our experiments attempt to shed light on some of the factors that might give rise to instances where frame accuracy and word error correlate. Some of the results are confirmation of intuitive or commonly known trends.

1. Introduction

In typical hybrid ANN-HMM (Artificial Neural Network- Hidden Markov Model) systems, an ANN is trained to estimate the posterior probability of subword-unit classes (e.g. phones) given a frame of acoustic features [1]. Frame accuracy (the percentage of frames where the maximum posterior corresponds to the correct class) is often used to measure the performance of the trained ANN. Where cross validation is employed, it is also used to measure the classification generality and to stop training. Our previous work on improving the acoustic modeling through discriminative training of the feature extraction as well as the ANN led to repeated cases where significant improvements in the frame accuracy led to a disappointing effect on word recognition [2]. This uncertain effect of an acoustic-modeling improvement has been witnessed by many ASR researchers. Unfortunately, it is difficult to analyze the precise relationship between the frame-level posteriors and word recognition. Such an analysis would further be specific to many aspects of the system, such as the HMM model architecture and parameters, the decoding algorithm, and the language model.

Arguably, a good frame classification accuracy is important for reasonable word recognition. We speculate, however, that the placements of the accurate frames has large influence on the ASR performance, and that not all frames have equal importance in the ASR system. To demonstrate this, we conducted a number of artificial experiments. Our approach was to perform numerous recognition tests with a controlled sequence of frame posteriors to obtain a sampling of the word error distribution for the system. Such an approach may reveal trends in the behaviour of the recognition system. Unlike a detailed sensitivity analysis, such a method can be repeated trivially should the system or system parameters change.

*Now with Voice Signal Technologies, Inc. (www.voicesignal.com)

2. Method

The method we used to perform controlled experiments was to artificially modify the frame classification rate of a probability stream from a data set prior to decoding. First a base sequence of class posterior probability estimates that had a relatively low frame accuracy was obtained from previous mis-matched reverberant environment tests. This probability sequence had a frame accuracy of 45% relative to the reference phonetic hand-transcription. We then corrected an additional 25% of the total number of frames (or 38% of the incorrect frames) to yield a total frame accuracy of 70%. Frames were randomly selected from a pool of all inaccurate frames and corrected by assigning a high posterior probability to the correct classes (from the reference transcription), while distributing the remaining probability mass equally among the remaining classes. Afterwords, word recognition was performed using the CHRONOS decoder [3] with monophone HMM states and with fixed decoding parameters. Tests were conducted with the development set of the NUMBERS corpus [4] which has a vocabulary size of 32 words. A summary of the original sequence is shown in the following table.

Number of utterances	1206
Number of frames	216518
Number of incorrect frames	118979 (55%)
Number of correct frames	97539 (45%)
WER (of 4673 words)	$\approx$ 40%
Number of frames to fix to achieve 70% frame acc.	54024 (25%) (38% of incorrect)

We chose to modify a relatively poor-performing sequence of probabilities rather than constructing a purely artificial one for practical reasons. We wished to start with probabilities that were generated from real features that would contain realistic posterior values, errors, and confusions. It would be non-trivial to construct a purely artificial sequence with these characteristics. In particular, the distribution of the probability mass among the non-correct classes would be difficult to do in a principled manner. Further, it is more convenient to correct the frames that were originally inaccurate than to corrupt correct frames in a realistic fashion.

3. Experiments

In all of the following, a total of 54024 of the 118979 misclassified frames were corrected to bring the frame accuracy to 70% of the total number of frames. In each of the experiments, the frames to be corrected were randomly chosen among either

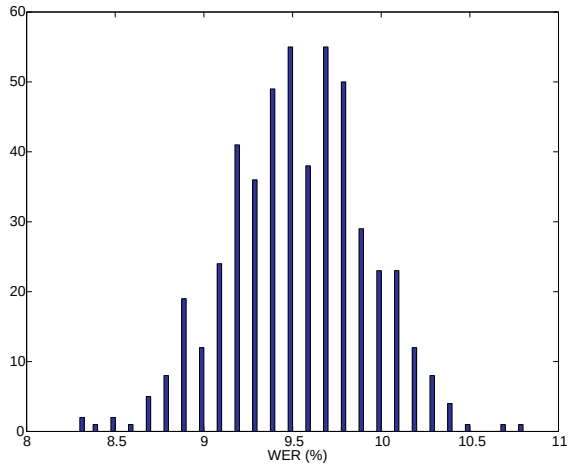

Figure 1: Histogram of WER for 500 recognition runs. In each run random incorrect frames were corrected to yield a frame accuracy of 70%.

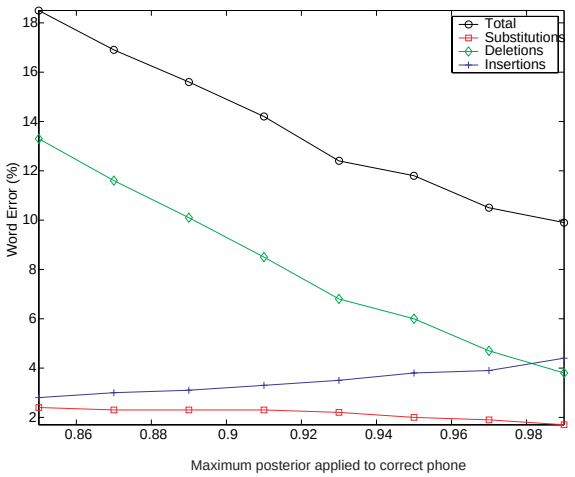

Figure 2: WER for one recognition run of randomly chosen corrected frames where the value of the posterior placed corrected frames was varied.

the total number of incorrect frames or a subset of frames that matched a given criterion. Random frames were selected by uniformly shuffling a list of the candidate frames and selecting a portion of them. Specific random seeds were assigned to permute the random numbers and to allow random sequences to be changed or duplicated or recovered.

3.1. Uniform Random Frame Correction

We ran 500 word-recognition experiments where the fixed number of corrected frames were randomly chosen among all of the incorrect frames. A different random frame sampling was chosen between recognition runs. The corrected frames were given a posterior probability of 0.99 in the correct class. A histogram of the resulting word error rates is shown in Figure 1. With a constant frame accuracy but a difference in selected correct frames, the resulting WER varied from 8.3% to 10.8%. Those runs with a WER higher than 10% or lower than 9% are significantly different from 9.5%.

Note that the original correct frames, 45% of the total frames, were the same for all runs. This test demonstrates that the placement of the correct frames can have a significant effect on the WER even though the total correct number of frames remained the same. The frame corrections were randomly chosen, equally among all incorrect frames. In subsequent tests, where certain frame types were corrected preferentially, WER scores sometimes varied by a much wider margin.

3.2. Posterior Value of Corrected Frames

In the previous tests, corrected frames had a high posterior of 0.99 assigned to the correct phone class with the remaining probability mass distributed equally among the rest of the phone classes. Frame accuracy, however, is a summary based upon the maximum posterior classification. The value of the maximum posterior can be much lower, as low as $\frac{1}{\#\text{phones}} + \epsilon$ while still being considered correct. The value has a direct bearing upon word recognition depending upon the probabilities associated with the surrounding frames. We conducted an additional test where the assigned corrected probability lowered from 0.99 to 0.85 in 0.02 decrements. Results from a single run using a fixed sequence of corrected frames is in Figure 2.

Varying the maximum posterior to something less or more ”confident” significantly altered the resulting WER. Even though each data point in Figure 2 is from a probability sequence with the exact same frame accuracy with the exact same frames classified correctly, the WER varied between 10% and 18%. This is not so difficult to believe since the decoded path must rely on the confidence of neighboring frames. Admittedly, the experiment is artificial and the pattern of frame probabilities is no longer ”natural.” The incorrect frames were fixed randomly with possibly many isolated among a group of incorrect frames. This is a possible shortcoming of the technique we have chosen to use here. However, correcting frames with a high posterior is necessary to force a new search path and overcome deficiencies in the surrounding frames. Correction with a high posterior allows us to observe indications of the importance of the placement of correct frames.

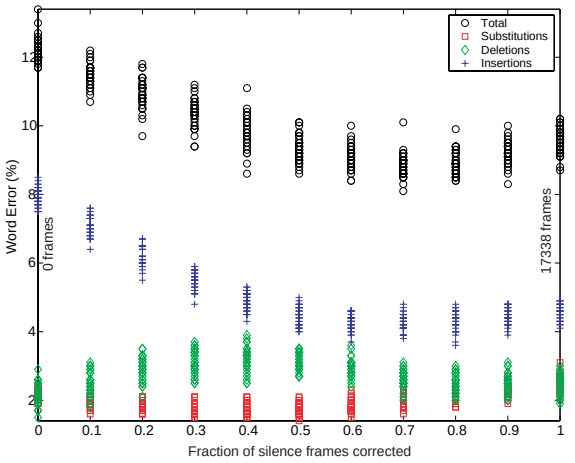

Figure 3: WER for 20 recognition runs with a varying proportion of corrected silence frames.

3.3. Corrected Silence Frames

Correctly determined locations of silence has an important function in segmentation, both of words and utterances. This next test makes a further distinction between the silence frames and the non-silence frames within the total number of incorrect

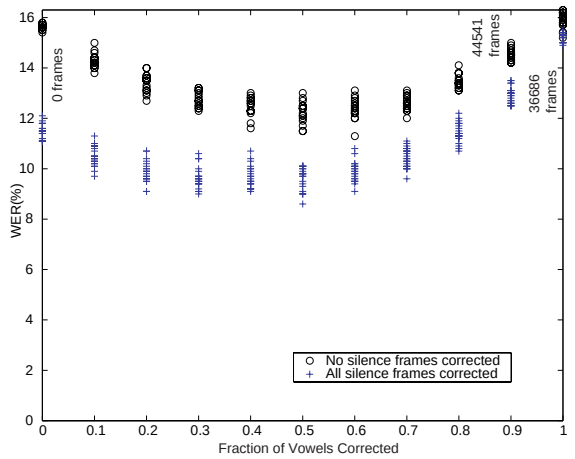

Figure 4: WER for 20 recognition runs with a varying proportion of vowel frames corrected. All silence frames were either corrected or left uncorrect independently.

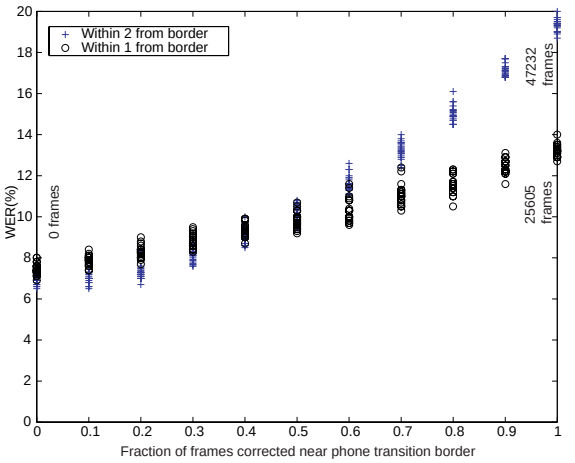

Figure 5: WER for 20 recognition runs with a varying proportion of the corrected frames that bordered phone transitions or were within 2 frames of the border transition.

frames. Proportions of the silence frames were corrected separately from the non-silence frames.

Silence frames incorrect	17338 (15%)
Non-silence frames incorrect	101641 (85%)

In Figure 3, the recognition tests were run with varying numbers of corrected silence frames ranging from no silence frames corrected to all of the silence frames corrected. All the while, the total frame accuracy was fixed at 70% of all frames. Thus, when more silence frames were corrected then fewer non-silence frames were corrected and vice versa. This was done 20 times with different selected frames. Again, corrected frames were given a posterior of 0.99 in the correct phone.

Figure 3 also displays the substitution, deletion, and insertion error subtypes. The number of insertions had the most prominence in the total word error, likely owing to restrictions silence places on word boundaries. As the number of corrected silence frames increased, the number of insertions went down. Past a certain point (70% of the silence frames), the number of substitutions began to rise, due to less non-silence frames being corrected. In these tests, the number of corrected silence frames and the WER are strongly and negatively correlated with a coefficient of -0.86. Further, silence constitutes only 15% of the incorrect frames, but makes a significant impact. Correct detection of silence is important for low WER.

3.4. Corrected Vowel Frames

These next two tests repeated the previous test except that frames corresponding to vowels were distinguished from the remaining phones. Vowels largely constitute the syllable nuclei. Therefore, these also tested to some degree the importance of syllable nuclei versus non-nuclei except that silence is a competing factor. The number of incorrect frames in the silence, vowel and non-vowel groups is shown below. 4.

Silence frames incorrect	17338 (15%)
Vowel frames incorrect	44541 (37%)
Non-vowel, non-silence frames incorrect	57100 (52%)

The WER results from 20 separate recognition runs are plotted in Figure 4. In the first set of data points, marked with '+', all of the silence frames were corrected with remaining incorrect frames portioned between vowel and non-vowel frames. The fixed silence phones reduced the number of allowed corrections so that only 36686 of the 44541 frames were candidates for correction. In the second set of data points, marked with 'o', none of the silence frames were corrected. With the silence phones corrected and somewhat removed from consideration, the fraction of vowel phones corrected has a correlation coefficient with WER of 0.56. Insertions was the principal error type in these test though substitutions seemed to follow the total WER best. With none of the silence frames corrected, the correlation coefficient between the fraction of vowels and the WER is 0.05, a very weak correlation. The insertions due to the uncorrected silence frames increased the WER level to between 12% and 16%. At this level it appears that a more or less equal proportion of corrected vowels and non-vowels is needed. There seems to be a balance between vowel and non-vowel phones such that some number of each is best. However, from the extreme ends (none or most vowels corrected) and from the correlation coefficients, it appears that correcting the consonants offers a slightly greater benefit, depending on silence accuracy.

3.5. Frames Bordering Phone Transitions

These two test examined the effect of incorrect frames near transitions from one phone to another in the reference transcription. In the first test, fractions of the number of incorrect frames that bordered phone transitions in the reference transcription were corrected. In the second test, incorrect frames that were within 2 frames from the transition were grouped and randomly corrected. Results from these tests are plotted in Figure 5. 20 recognition runs with different selected random frames were performed in each of the tests.

Incorrect frames bordering transitions	25605 (22%)
Incorrect frames not bordering transitions	93374 (78%)
Incorrect frames within 2 frames from border transitions	47223 (40%)
Incorrect frames not within 2 frames from border transitions	71756 (60%)

The fraction of corrected frames that border phone transitions is strongly correlated with WER with a coefficient of 0.97. This is true for both tests. It is interesting to see the WER for the

second set of tests (incorrect frames within 2 frames of the border) rise from about 7% to almost 20%. All experiments have exactly the same frame accuracy. To the extreme right in the plot, all of the transition-bordering frames were corrected with relatively few (6801) of the remaining frames corrected. To the extreme left in the plot, only non-transition-bordering frames were corrected; incorrect transition-bordering frames were left unaltered. This resulted in the best group of WER. The WER when assigning corrected frames away from the transition borders is lower than the average WER from a uniform random assignment (Figure 1). From these tests, it seems that corrections that are nearer the centers of the phones are more important than near the boundaries, though other factors contribute. These tests used hand-transcribed phonetic transcriptions as the reference for both training the probability stream and classification summaries. The tests therefore rely on accurate phonetic segmentation as well as identity. Precise placement of transitions between phones can be dubious for many pairs of phones. It is therefore encouraging that precise classification at the boundaries may not be necessary.

3.5.1. Removing Transition Frames from Training Data

We conducted additional tests where our ANN probability estimator was trained with and without the transition-bordering frames. We used a three layer Multi-Layer Perceptron (MLP) with a single frame of acousitic PLP features as input and 400 hidden units. The MLP was trained with clean data from the NUMBERS corpus and tested with clean data and data with artificially added factory noise at 10dB SNR. Results from the test are in the following table.

Test	WER(%)	Facc(%)	bFacc(%)
clean	7.3	71.03	74.34
clean no border	7.7	70.20	74.31
factory	15.5	57.16	60.42
factory no border	15.0	57.12	60.85

Here, “Facc” denotes frame accuracy while “bFacc” denotes frame accuracy of only the non-transition-bordering frames. There was no significant change in the WER and frame accuracy when training the MLP with and without transition-bordering frames. Training without bordering frames unfortunately did not increase the accuracy of the non-bordering frames. Further, doing so did not noticeably increase the average posterior values of the correct non-bordering frames, which would have resulted in improvements. However, it did produce nearly equivalent WER results using 80% of the training data.

4. Discussion

Word recognition error depends upon the accurate classification of the frame probabilities, the locations of the errors and the frame posterior values. A thorough investigation of the relationship between frame accuracy and word recognition would require a more detailed sensitivity analysis of the decoding system and the models. Such an analysis is non-trivial to construct and is dependent upon the decoding algorithm and its parameters. Though less ideal, the random selection approach conducted here is a general empirical method that is independent of the specific decoder and can yield some indication as to factors that are important for word recognition. A random sampling of corrected frames gives rise to a distribution of corresponding word error rates despite equal overall frame accuracy. Varying the proportion of some types of frame errors can yield results that vary in a systematic fashion. Depending upon the proportion of errors, the resulting WER can vary by a significant amount.

With these complications, the frame accuracy is not necessarily a proper measure when comparing the probabilities of two or more acoustic sequences. Since the value of the maximum posterior can have a strong effect on WER, we also considered a frame accuracy weighted by the posterior values for the correct class and an average of the posteriors for the correct classes. Computed measures were, however, only weakly correlated with WER, with a coefficient of -0.10. Additional weighting could be included if it is determined that certain types of frames are more important than others in the resulting decoding. For example, the silence frames are relatively important whereas the transition bordering frames may not be. Naturally, further tests are needed for a better picture.

5. Conclusion

Our tests examined to some degree the location of frame errors depending on criteria such as silence, vowel and phone transition. They were shown to have some systematic effects on the word error distribution. Further tests associated with model states can be conducted with other decoders that provide decoding lattice information. Future analysis may be combined and compared to related work by Chang et. al. who analyzed frame errors relative to phone position with-in words and syllables [5] and Greenberg et. al. who conducted ASR diagnostic evaluations with respect to many acoustic, linguistic and speaker characteristics [6]. Results from future diagnostics may aid in selecting and training front-end acoustic modeling in a manner better suited to the overall goal of word recognition.

6. Acknowledgements

We would like to thank Nelson Morgan and ICSI for supporting this work. This work was funded by NSF, grant number (NSF)-IRI-9712579.

7. References

[1] Hervé Bourlard and Nelson Morgan, *Connectionist Speech Recognition- A Hybrid Approach*, Kluwer Academic Press, 1994.

[2] Michael L. Shire and Barry Y. Chen, “On data-derived temporal processing in speech feature extraction,” in *International Conference on Spoken Language Processing*, Beijing, China, October 2000, vol. 3, pp. 71–4.

[3] Tony Robinson and James Christie, “Time-first search for large vocabulary speech recognition,” in *International Conference on Acoustics Speech and Signal Processing*, Seattle, Washington, May 1998, IEEE, vol. 2, pp. 829–32.

[4] Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute, “Numbers corpus, release 1.0,” 1995.

[5] Shuangyu Chang, Lokendra Shastri, and Steven Greenberg, “Automatic phonetic transcription of spontaneous speech (American English),” in *International Conference on Acoustics Speech and Signal Processing*, Beijing, China, October 2000, vol. 4, pp. 330–3.

[6] Steven Greenberg, Shuangyu Chang, and Joy Hollenback, “An introduction to the diagnostic evaluation of Switchboard-corpus automatic speech recognition systems,” in *Proceedings of the NIST Speech Transcription Workshop*, College Park, Maryland, May 2000.