APPENDIX A. SIGNIFICANCE TEST

The main goal of this project was to facilitate the comparison of front-ends on a speech recognition task. To do this, we need to be able to say, definitively, that one front-end outperforms another. A first-cut analysis might say that if one has a lower WER then it must be better. However, there are many parts of an ASR system that are independent of the front-end (e.g. the number of processors used for parallel training) but that could introduce noise into the recognition results. There is, thus, a need for a statistical measure which can give us confidence that one system is truly better than another. Statistical significance tests were employed in this work to fill this need.

The objective of statistical testing is to make inferences about a population given a finite sample set from the population. Since we lack complete knowledge of the population, we have a need to measure the uncertainty in our inferences. In statistics, this uncertainty is often referred as the probability of making an error (in our inference). Statistical testing provides a means quantitateive evaluation and control over the probability of making an error.

Statistical testing is often posed as a problem of hypothesis testing. A statistical hypothesis is a speculation about the population's behavior or is some established theory stated in terms of population parameters such as means and variances. Data is collected during an experiment, or set of experiments, and is assumed to be drawn according to the population distribution. Statistical tests provide a means for making judgements as to the truth of the statistical hypothesis based on the sample data. A control parameter known as the *significance level* is used to control the probability of making an error in the inference. The significance level measures our confidence in rejecting the hypothesis. The lower our confidence of rejection is, the more likely it is that our inference is true.

A common task in hypothesis testing is to compare statistics computed over samples of two distributions to determine how likely it is that the two distributions are equivalent. For example, we may want to compare the means and variances of two sampled distributions, each of which is assumed Gaussian with with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. Consider the case for comparing the means of the two populations. We begin by forming the null hypothesis that the two means are equivalent:

Null Hypothesis H0: $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$

Alternate Hypothesis H1: $\mu_1 \neq \mu_2$ or $|\mu_1 - \mu_2| > 0$

We randomly select n_1 samples from the first population and then draw n_2 samples independently from the second population. The difference between the two sample means $\overline{y_1} - \overline{y_2}$ is an unbiased point estimate of the difference of the true population means $\mu_1 - \mu_2$. According to the linear function of the random variables, the sampling distribution of statistic $\overline{y_1} - \overline{y_2}$ is a normal distribution with a mean of $\mu_1 - \mu_2$ and variance of $\sigma_1^2 / n_1 + \sigma_2^2 / n_2$. Hence, the test statistic or z-statistic is given by

$$Z = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}}$$

This test statistic's distribution can be approximated as a standard normal distribution shown in Figure 47. A single right tailed test can be used to reject the null hypothesis when $Z = z_p$ at a significance level of p. The rejection region or the probability of falsely rejecting the true null hypothesis (Type I error) lies in the region from z_p to infinity. This region as shown as yellow region in the Figure 47.

The problem in our work is to specify an upper limit for performance (WER) for which a new design of a frontend would be considered to be statistically significantly better than the baseline. *Significance for proportions* is suitable for such needs since WER is defined as a proportion. This leads to the same form as the z-test. The assumption is made that two experiments (baseline and new front-end system) each consisting of N independent trials are run. To satisfy the independence assumption, it is necessary to consider each trial as the number of errors for each utterance in the corpus. This requires the assumption that the utterances in the corpus were independent of each other. For example, the utterances in the corpus should not be derived from discussions where one utterance is a response to another. We can not use the words in the corpus as trials since we know that n-gram language models dictates that consecutive words are not independent of each other.

If, in our experiment, the first experiment resulted in y1 trials in error while the second experiment resulted in y2 trials in error, we can estimate the word error rates, p1 and p2 from a sample of size N in the sample population, $\hat{p}1 = y1/N$ and $\hat{p}2 = y2/N$. Since the aim is to determine if the second experiment WER, p2, is significantly better than the first experiment WER, p1, given N trials for each experiment, we consider the difference of the word error rates (proportions) to be zero as the null hypothesis, H0:

Null Hypothesis H0: $p1 = p2$ or $p1 - p2 = 0$

Alternate Hypothesis H1: $p1 \neq p2$ or $|p1 - p2| > 0$.

To prove that the second experiment p2 is significantly better than the first experiment, we need to reject H0 at a given significance level. The normalized z-statistic for this test is given as:

$$Z = \frac{(\hat{p}1 - \hat{p}2)}{\sqrt{\left(\frac{\hat{p}1(1-\hat{p}1)}{N}\right) + \left(\frac{\hat{p}2(1-\hat{p}2)}{N}\right)}}$$

The assumption for this test is that according to the Central Limit Theorem, the distribution of this z-statistic is approximately normal given a large sample size. The single-tailed significance test is used to reject or fail to reject the null hypothesis. Note that the variance of $p1 - p2$ is estimated in the denominator of the equation above.

The baseline performance, 15.4% WER, of the short test set 1 consisting of 166 eval utterances and training condition I (8 KHz sampling rate, utterance detection, clean training condition and compressed features) was taken as the reference for the current ETSI frontend. The aim was to specify the upper-limit of the WER that would be accepted as significantly bettter than the baseline. With, N=166, p1=0.154 and significance level of 1% (p=0.001), we iterate over decreasing value of p2 starting from 0.154 until the null hypothesis is rejected. It can be shown that when p2 reaches an error rate of 0.073 or 7.3% WER, the z-statistic is given by

$$Z = \frac{(0.154 - 0.073)}{\sqrt{\left(\frac{0.154(1-0.154)}{166}\right) + \left(\frac{0.073(1-0.073)}{166}\right)}} = 2.3455476$$

Since $z_{0.01} = 2.3263479 > 2.3455476$ and $z_{0.02} = 2.0537489 < (2.3455476)$, we reject the null hypothesis at the 1% significance level. Similarly it can be shown that at 10% significance level the value of p2 is 0.106 or 10.6%. Thus, 7.3% WER and 10.6% WER specify the upper bound on significant results for the 1% and 10% significance levels, respectively.

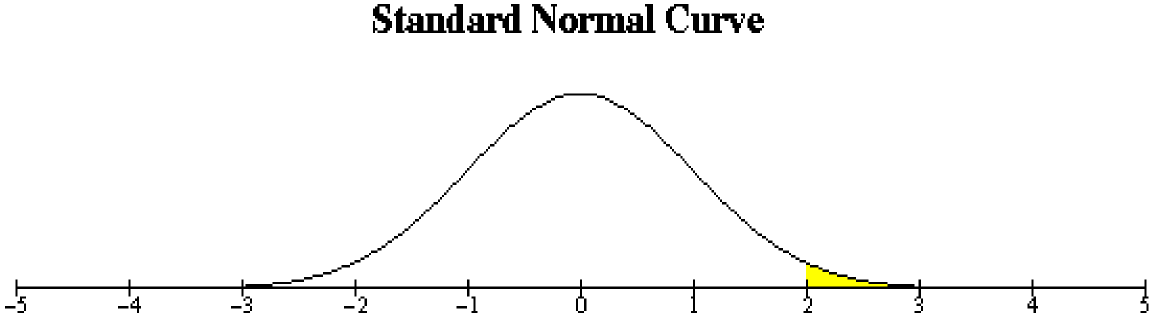

Figure 47. Standard Normal Distribution. Rejection region is marked as yellow.