Improving Perplexity Measures To Incorporate Acoustic Confusability

Amit Anil Nanavati *Nitendra Rajput*

IBM India Research Lab
Block 1, IIT Campus,
Hauz Khas, New Delhi, 110016. India
{namit,rnitendra}@in.ibm.com

Abstract

Traditionally, Perplexity has been used as a measure of language model performance to predict its goodness in a speech recognition system. However this measure does not take into account the acoustic confusability between words in the language model. In this paper, we introduce *Equivocality* – modification of the perplexity measure for it to incorporate the acoustic features of words in a language. This gives an improved measuring criterion that matches much better with the recognition results than conventional Perplexity measure. The *acoustic distance* is used as a feature to represent the acoustic characteristic of the language model. This distance is measurable only with the acoustic model parameters and does not require any experimentation. We derive the *Equivocality* measure and calculate it for a set of grammars. Speech recognition experiments further justify the appropriateness of using *Equivocality* over Perplexity.

Index Terms: perplexity, language model, acoustic distance, word error rate.

1. Introduction

Language modeling has been used in speech recognition systems to reduce the acoustic search space. For constrained tasks such as conversational systems, grammars are typically used to represent the language model for each user input. The perplexity of these grammars is (supposedly) a measure of the difficulty of recognising the grammars by a speech recognition system. Though the true quality of a language model or a grammar can be evaluated only by performing a speech recognition experiment, but performing this experiment may not always be possible. Therefore Perplexity is used to evaluate the grammars and language models in a isolated manner. It provides an information theoretic measure for predictability of a language syntax. Although perplexity does not use any phonetic character of the words in a language, this has been the most commonly used measure for comparing language models and grammars for different tasks.

The correlation between grammar perplexity and the Word Error Rate (WER) of a speech recogniser is shown in [4]. However there are cases when this correlation fails. In [2], a technique of calculating the WER without requiring a speech recognition system is presented. Modifications to the perplexity measurement technique is also presented in [1]. Here the authors adapt the Shannon game for evaluation of language models.

In this paper, we demonstrate that the perplexity measure of speech recognition grammars do not have a close correlation with the corresponding WER. We establish this by performing relevant experiments in Section 2. We then formulate the measurement of acoustic characteristic of the corresponding grammar. If a grammar has words that are acoustically similar (confusing), then that grammar is expected to give a higher WER compared to an exactly similar structured grammar, but with different sounding words. We define the acoustic distance between words and present a measure of the confusability of the grammar in the acoustic space. We then present a formulation that incorporates these acoustic distance between words in a information-theoretic framework. Section 3 presents this solution. We substantiate the *Equivocality* measure using several test grammars and by performing speech recognition experiments in Section 4. The improved correlation of *Equivocality* with WER verifies the mathematical formulation. The paper concludes with alternative ways for defining *Equivocality* in Section 5.

2. BACKGROUND AND MOTIVATION

An information theoretic view of the perplexity shows that the entropy or the information per word that is associated with a word sequence $(w_1, w_2, ..., w_n)$ is defined as

$$H = -\frac{1}{n}\left[\log_2 P\left(w_1, w_2, ..., w_n\right)\right] \quad (1)$$

When a word sequence is less likely, the information content in that sequence is high, as is seen from equation 1. The term perplexity PP is related to the entropy as seen below:

$$PP = 2^H \quad (2)$$

For a grammar g that has 10 isolated words, with all being equally likely, the perplexity is $PP(g_1) = 2^{-\left(\log \frac{1}{10}\right)} = 10$.

Since the measure for perplexity does not depend on the pronunciation (acoustic proximity) of these words, it will remain the same for all 10–word grammars that have all the words as equally likely. We use three grammars for conducting a speech recognition experiment. These grammars are shown in Table 1. Though the perplexity of these three grammars are the same, the WER is very different as seen in the table.

Table 1 implies that two grammars with the same perplexity can have different WER. But the absence of correlation between the perplexity and WER is highlighted by the grammars g_1, g_2 and g_3. In this example, the perplexity is not able to capture the acoustic closeness of words that exist in a grammar. Therefore, grammars that have confusing words such as *bait, bat* and *bet* have a higher WER.

The above experiments were performed on an IBM ViaVoice speech recogniser. The details of this recogniser are presented in

Table 1: Perplexity and WER for isolated word grammars

Grammar	Perplexity	WER
$g_1 = zero\|one\|two\|three\|four$ $\|five\|six\|seven\|eight\|nine$	10	5
$g_2 = john\|tom\|sam\|bon\|ron\|$ $\|susan\|sharon\|carol\|laura\|sarah$	10	7
$g_3 = bit\|bite\|boot\|bait\|bat$ $\|bet\|beat\|boat\|burt\|bart$	10	9

Section 4. These experiments clearly highlight the fact that perplexity is alone not a good measure of the performance of the grammar in a speech recogniser. Additionally, the experiments suggest that a notion of acoustic distance between words in a grammar should provide more insights into the performance of the grammar in a speech recognition system. This forms the basis of this paper. In the next section we present the use of acoustic closeness in the perplexity measure.

3. OUR SOLUTION

With the examples in Section 2, it is clear that a measure that captures acoustic closeness between words along with the perplexity should be more correlated with the WER. Perplexity is a measure which is based on the amount of information that is received when a particular work sequence *occurs* in the language. We model Equivocality – it captures the amount of information that is received when a particular word sequence is *recognised*. Therefore we show that Equivocality provides a better estimate of the grammar performance in a speech recogniser. In this section, we first define a measure of acoustic confusability between words. Later on we model Equivocality that uses the acoustic distance between words in a information theoretic model.

3.1. Acoustic distance between words

Word confusability has been measured in the past through the acoustic model and the phonetic content of each word [6, 3]. The confusability metrics method has been used in [5]. We use the confusability measures (similar to [6]) to formulate the measure for distance between two words.

At a higher level, the distance between two words can be extracted by using a string matching algorithm on the phonetic spelling of two words. For example, the distance between the two words *bon* and *ron* will be calculated as follows:

bon	B	OW	N
ron	R	OW	N
	1	0	0

Thus the distance between these two words is 1 unit. On the other hand, for words of unequal length, an insertion, deletion and a substitution can be considered as of unit length distance the complete distance between words can be calculated by counting the number of such operations. These techniques are used in calculating distance of one string from other and are well studied in the literature [7]. For example, the distance between *sam* and *susan* can be obtained as follows:

sam	S	_	_	AE	M
susan	S	UH	Z	AE	N
	0	1	1	0	1

The distance is thus 3. However it can be believed that the distance between N and M is far less than the distance between an insertion, i.e. between _ and UH. Thus a distance metric can be designed such that it has different weights to an insertion/deletion versus a substitution. Moreover, not all substitutions may have the same weight. If the Acoustic Model for a particular language is available, these distances can be calculated by measuring the distance between gaussians of each phone as in [6].

In general, for any two words w_1 and w_2, the distance can be measured by:

$$d(w_1, w_2) = \sum_{i=1}^{P} d_p(w_1^i, w_2^i) \tag{3}$$

where $d_p(w_1^i, w_2^i)$ is the distance between the hidden markov models of the two phonemes w_1^i and w_2^i and P is the total number of phones in each word, after insertions. This distance can be normalised by the units of measurements that are based on the space in which these HMMs have been trained. Different ways of calculating distance between the HMMs are presented in [6] and we will not go into details of this measure. We measure the distance between the two techniques – one that assigns equal weight to all insertion and substitutions and the other that uses the distance measure in equation 3.

3.2. Equivocality

The perplexity measure is based on the amount of self-information associated with a word sequence. In order to find a measure that relates more closely with the WER in a speech recognition experiment, we model *Equivocality* in this section.

As mentioned in equation 1, the self-information is measured by the probability of word sequences. If the words are independent, this can be written as

$$H = -\left[\sum_{i=1}^{L} P_o(w_i) \log P_o(w_i)\right] \tag{4}$$

where $P_o(w_1), P_o(w_2, ..., P_o(w_n)$ are the original probabilities of the *occurrence* of the words in a given grammar. So

$$\sum_{i=1}^{n} P_o(w_i) = 1 \tag{5}$$

However, in order to be effective for speech recognition, each word that can *occur* should also be correctly *recognised* by the speech recogniser. From this point of view, the probability of occurrence for each word can be subdivided into probability of recognition and of mis-recognition.

$$P_o(w_i) = P_r(w_i) + P_{mr}(w_i) \tag{6}$$

Therefore

$$\sum_{i=1}^{n} P_o(w_i) = \sum_{i=1}^{n} P_r(w_i) + \sum_{i=1}^{n} P_{mr}(w_i) \tag{7}$$

Using 6 and 7, we have

$$\sum_{i=1}^{n} P_r(w_i) = 1 - M \tag{8}$$

where $P_r(w_1), P_r(w_2, ..., P_r(w_n)$ are the probability of *recognition* of these words and M represents the total probability of mis-recognition by the speech recogniser. The value of M is dependent on the acoustic confusability of the words in the grammar and also on the speech recogniser. Since the speech recogniser can be thought of as a constant performance system, the only variable is the confusability of the words in the grammar. We use the measure of acoustic distance between words to formulate M. Further, as seen from equation 8, the set of $P_r(w_i)$ probabilities are not a complete probability set, but it is a representative of a significant subset of the complete set $P_o(w_i)$ that better correlates a grammar with its speech recognition performance. This will be further validated in Section 4.

In order to find the value for $P_r(w_i)$'s, we calculate the acoustic distance of a word w_i to all its alternatives, that is, words that could potentially occur at the same position (linguistically) as w_i.

Suppose there are n alternative words in the grammar at a particular linguistic position. Then, the average distance:

$$d_{avg}(w_i) = \frac{\sum_{j \neq i}^{n} d(w_i, w_j)}{(n-1)} \tag{9}$$

where $d(w_i, w_j)$ is calculated as shown in equation 3. The average distance measure $d_{avg}(w_i)$ captures of how distant (distinguishable) w_i is from the rest of the words in the grammar. If there is a single word in the vocabulary, then $d_{avg}(w_1) = \infty$. If there are at least two distinct words, then $d_{avg}(w_i) > 0$ regardless of how close the words might be. Therefore, $0 < d_{avg}(w_i) \leq \infty$.

We now quantify the probability of mis-recognition of word w_i. Intuitively, the probability of mis-recognition of a word is proportional to the probability its occurrence. The more frequently a word occurs, the more likely it is to be mis-recognised (as well as recognised). If two words occur with the same probability, the word which has acoustically closer neighbours is more easily confused and harder to recognise. We capture this simple intuition as follows:

$$P_{mr}(w_i) = f\left(P_o(w_i), d_{avg}(w_i)\right) \tag{10}$$

Since the *proximity* decreases with the increase of this distance and vice versa, this inverse relationship results in:

$$P_{mr}(w_i) = \frac{P_o(w_i)}{d_{avg}(w_i)} \tag{11}$$

This measures the probability of mis-recognising a single word, w_i. Summing up the probabilities of mis-recognition of all words, we have M:

$$M = \sum_{i=1}^{n} P_{mr}(w_i) = \sum_{i=1}^{n} \frac{P_o(w_i)}{d_{avg}(w_i)} \tag{12}$$

Therefore M is the harmonic mean of the distances weighted by their occurrence probabilities. This inverse of distance clearly indicates *proximity*, so the likelihood that *none* of the words are recognised is given by the summation of the proximities of words weighted by their occurrence. Thus, M also quantifies the probability that *none* of the words is recognised, and provides an explanation of why $\sum_i P_r(w_i) < 1$. This is a fundamental difference between the probabilities of occurrence and those of recognition: it is impossible that none of the words occur, while it is possible that none of them may be recognised. In fact, M characterises the complete layout of the words in the acoustic space and also integrates the linguistic and acoustic factors in a simple, intuitive equation. As expressed in equation 8, M determines the limits on how much can be recognised.

To find $P_r(w_i)$, we notice from equation 8 and 12 that

$$\sum_{i=1}^{n} P_r(w_i) = \sum_{i=1}^{n} P_o(w_i) - M \tag{13}$$

$$\implies P_r(w_i) = P_o(w_i) - P_{mr}(w_i) \tag{14}$$

Using this, we find the values for P_r by modifying equation 14

$$P_r(w_i) = P_o(w_i)\left(1 - \frac{1}{d_{avg}(w_i)}\right) \tag{15}$$

To accommodate for values of $0 < d_{avg}(w_i) < 1$, we modify this to

$$P_r(w_i) = P_o(w_i)\left(1 - \frac{\epsilon}{d_{avg}(w_i) + \epsilon}\right), \quad \epsilon > 0 \tag{16}$$

Therefore,

$$P_r(w_i) = P_o(w_i)\left(\frac{d_{avg}(w_i)}{d_{avg}(w_i) + \epsilon}\right) \tag{17}$$

The role of ϵ in the above equations is two-fold. Firstly, it ensures that $P_r(w_i)$ never exceeds $P_o(w_i)$, when $0 < d_{avg}(w_i) < 1$ and secondly, as a parameter to control the effect of the acoustic space. It provides a handle to weight the acoustic part with the linguistic part in the joint estimation. These set of probabilities $(P_r(w_1), P_r(w_2), ..., P_r(w_n))$ can be interpreted as not being the probability of occurrence but the probability of recognising that occurrence. Therefore using these probability set in the calculation of entropy will be a better measurement of the grammar or the language-model. The *Equivocality*, thus, can be defined using these set of probabilities as

$$\bar{H} = -\frac{1}{n}\left[\log_2 P_r(w_1, w_2, ...w_3)\right] \tag{18}$$

and

$$EV = 2^{\bar{H}} \tag{19}$$

Since the P_r probabilities are always lower than the P probabilities (by a cumulative difference of M), EV will always be higher than PP. Moreover, this difference between EV and PP depends on the acoustic goodness of the words in the grammar. If $M = 0$, then $EV = PP$, thus implying that the information content in recognising an occurrence is same as the information content in the occurrence since the recognition is a certain event.

The utility of this formulation is 3-fold. First, it unifies linguistic and acoustic consideration for characterising a grammar. Second, it provides a more accurate measure for grammar performance. Third, that it can be calculated analytically, without performing experiments for every new grammar.

We now calculate the distances $(d_{avg}(w_i))$ and probabilities of recognition $(P_r(w_i))$ using equations 3, 15 for some example grammars. Then using equation 18, 19, we calculate the EV for those grammars and observe their correlation with the WER experiment results in the next section.

Table 2: *M* and *Equivocality* for grammars

Grammar	M	Equivocality
g_1	0.212	5
g_2	0.377	7
g_3	0.563	9

Table 3: *Perplexity*, *Equivocality* and WER relationships

Grammar	Perplexity	Equivocality	WER
g_1	10	5.83	5
g_2	10	6.73	7
g_3	10	7.83	9

4. EXPERIMENTS AND RESULTS

We use the same grammars of Table 1 to find the distance between their words and to evaluate the *Equivocality* of each grammar. The average distance of these words are calculated using the distance metric formulated in equation 3. We used the Acoustic Model of an Indian English speech recognition system to generate the distance between words. This model was based on a set of 65 phones and had two silence phones *X:* intra-phrase-silence and *D$*: intra-word-silence. Each phone was modelled by an HMM. The HMMs are tri-state left-to-right models. The observation and transition probabilities have been trained on a data of about 500 hours of continuous speech. In order to improve the acoustic space, the observations are further divided into context-dependent phones, which we call as leaves. Thus, the observations are modelled for a total of 3560 leaves. Further, since the amount of data and its variations is too large, each leaf is modeled by a set of mixture gaussians. In all, corresponding to 3650 leaves, there were 105926 mixture gaussians.

To perform speech recognition experiments, a small test was conducted using a couple of speakers with all the possible utterances in the three grammars. Since the WER values are illustrative of the grammar recognition difficulty, this number is enough to provide that effect.

The distance between every pair of words has been calculated by using the mean, variance and weight of these gaussians. For a given pair of words, their phone sequence is extracted from the phone vocabulary. Then insertions and substitutions are used to generate the two phone strings such that they have the least edit-distance [7]. Based on the phone-pairs thus obtained, the appropriate gaussians are selected and their mean values are used to generate the distance $d_{avg}(w_i)$ for each w_i. This exercise is repeated for all words in the grammar.

The M of the three grammars is calculated using equation 12. Once the distances $d_{avg}(w_i)$ are calculated, the values for $P_r(w_i)$ are generated. Then we generated the *Equivocality* numbers for each of these grammars. The value of *M* and *Equivocality* of the three grammars in Table 1 are shown in Table 2. The corresponding Table 3 clearly shows that *Equivocality* is able to distinguish between grammars that have the same occurrence probability $P_o(w_i)$ on the basis of their acoustic properties. Thus the increasing WER observed among these grammars corresponds with the *Equivocality* numbers for these grammars.

Though the illustrative grammars have been simplified to have equal probability for all words, the concepts work for any types of complex grammars. The illustration is intuitive to follow if simple grammars are used.

5. CONCLUSION AND FUTURE WORK

We introduced a new measure, *Equivocality*, that captures the linguistic as well as acoustic properties of words in a grammar. While Perplexity captures the linguistic model in terms of word *occurrence*, this is a poor indicator of the *recognisability* of a grammar, since it does not account for acoustics. The concept of acoustic distance between words have been applied to find the acoustic similarity between words in a grammar. We use acoustic distance along with linguistic probabilities to get a unified measure. The utility of this formulation is 3-fold. First, it unifies linguistic and acoustic consideration for characterising a grammar. Second, it provides a more accurate measure for grammar performance. Third, that it can be calculated analytically, without performing experiments for every new grammar.

For our experiments, we calculated *Equivocality* for a set of three grammars and found that the observed WER corresponds better with the *Equivocality* measure than with the standard *Perplexity* measure. Since measuring of *Equivocality* does not require any speech input, this can still be done given a grammar. The definitions hold for any Language Model or grammar.

The analysis for calculating the total mis-recognition yielded M to be the harmonic mean of the distances weighted by their occurrence probabilities. M rather neatly captures the interplay between the linguistic and acoustic factors, and in fact, characterises the complete layout of the words in the acoustic space. M alo determines the limits on how much can be recognised.

6. References

[1] F. Bimbot, M. El-beze, and M. Jardino. An alternative scheme for perplexity estimation. In *IEEE ICASSP*, volume II, pages 1483 – 1486, April 1997.

[2] S. Chen, D. Beeferman, and R. Rosenfeld. Evaluation metrics for language models. In *Proceedings of the Broadcast News Transcription and Understanding Workshop, sponsored by DARPA*, pages 275 – 280, 1998.

[3] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, and M. Parker. Automatic speech recognition with sparse training data for dysarthric speakers. In *Eurospeech*, pages 1189 – 1192, Geneva, Switzerland, September 2003.

[4] F. Jelinek. Self-organised language modeling for speech recognition. In *Readings in speech recognition, San Mateo, Morgan Kaufmann,*, pages 450 – 506, 1990.

[5] J. Peillon, J. Hernando, and A. Bramoulle. Word confusability prediction in automatic speech recognition. In *ICSLP*, pages 1489 – 1492, Jeju Island, Korea, October 2004.

[6] B. Tan, Y. Gu, and T. T. Word confusability measures for vocabulary selection in speech recognition. In *IEEE ASRU*, pages 185–188, Keystone, USA, December 1999.

[7] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, 1974.