IDIAP RESEARCH REPORT

# On the Use of Information Retrieval Measures for Speech Recognition Evaluation

Iain McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, Hervé Bourlard.

IDIAP–RR 04-73

March 2005

**IDIAP Research Institute** **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-73

# On the Use of Information Retrieval Measures for Speech Recognition Evaluation

Iain McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn,
Pierre Wellner, Hervé Bourlard.

March 2005

**Abstract.** This paper discusses the evaluation of automatic speech recognition (ASR) systems developed for practical applications, suggesting a set of criteria for application-oriented performance measures. The commonly used word error rate (WER), which poses ASR evaluation as a string editing process, is shown to have a number of limitations with respect to these criteria, motivating alternative or additional measures. This paper suggests that posing speech recognition evaluation as an information retrieval problem, where each word is one unit of information, offers a flexible framework for application-oriented performance analysis based on the concepts of recall and precision.

# 1 Introduction

While use of the word error rate (WER) as a common measure has served to advance speech recognition research in recent times, there is currently interest in the speech recognition community to consider evaluation measures that allow a more thorough analysis of system performance, and yield clearer interpretation, particularly in terms of end application usability.

Before defining an evaluation framework, it is necessary to consider the properties that it ideally should have. First, while it is necessary to analyse performance with respect to end usability, evaluation measures must still fundamentally be *direct* measures of the ASR component.

Second, the measure should be calculated in an *objective*, automated manner. Measures based on subjective, resource-intensive, and application-dependent user trials would tend to impede scientific comparison, as well as retard and splinter research.

Third, the measure must be clearly *interpretable* in the light of end application usability. The absolute value of the measure must have an intuitive relationship to system performance, and relative differences between measures should give a fair indication of their relative merits.

Finally, while the evaluation measure should be general, it should be *modular* to allow thorough application-dependent analysis. Different end applications place different relative importance on particular words and associate different costs with different types of errors. For instance, consider a application in which an alarm is generated if one of a set of keywords is spoken. In this case, only a small subset of all spoken words are important, and the balance between false alarms and missed alarms will depend on the relative costs of reacting to, or missing, the alarm. In a spoken document retrieval system, the relative importance of each word will depend on the information it carries with respect to the particular application context, and usability may be hampered more by missing information than by erroneous insertions. As a final example, in a dictation application, all words may be considered to be of equal importance, and missed words be just as costly as falsely inserted words. Any application-oriented evaluation framework must therefore allow the performance to be analysed in terms of individual words or particular types of errors and, where single value measures are required, these must allow the relative importance of words and costs of error types to be configured.

This paper discusses the issue of application-oriented evaluation of speech recognition systems. Summarising the above, these evaluation measures should be *direct*, *objective*, *interpretable* and *modular*. In this context, we discuss limitations of the word error rate, and show that the information retrieval concepts of *recall* and *precision* can meet each of these criteria.

# 2 Background and Related Work

## 2.1 The Word Error Rate: Definitions and Limitations

The word error rate, commonly used to evaluate ASR systems, is derived from the *Levenshtein distance*, or *edit distance*. The edit distance between two strings is the minimum number (or weighted sum) of insertions, deletions and substitutions required to transform one string into the other [1]. The WER is the edit distance between a reference word sequence and its automatic transcription, normalised by the length of the reference word sequence. This normalisation is applied to allow comparison between different systems on different tasks, as the magnitude of the edit distance depends on the string length.

We define: $N_r$ as the total words in the reference transcription, $N_a$ as the total words in the automatic transcription, $S$ as the number of substituted words in the automatic transcription, $D$ as the number of words from the reference deleted in the automatic transcription, $I$ as the number of words inserted in the automatic transcription not appearing in the reference, and $H$ as the number of correctly recognised words. The word error rate is defined as:

$$WER = \frac{S + D + I}{N_r}. \tag{1}$$

While this measure is most commonly used as an error rate, it is also often quoted as the *word recognition rate*,

$$WRR = 1 - WER = \frac{H - I}{N_r}, \tag{2}$$

noting that $H = N_r - (S + D)$. Sometimes, particularly for isolated word recognition systems, the *word correct rate* is used as a performance measure for speech recognition. It does not consider insertion errors, and is defined as

$$WCR = \frac{H}{N_r}. \tag{3}$$

In terms of the criteria suggested in the introduction, it is clear that the word recognition rate is a *direct* and *objective* measure of speech recognition performance. However, as we will explain in the following, the WER has a number of practical disadvantages which mean it is sometimes not easily *interpretable* in terms of application usability, and also render it an unsuitable framework for decomposing the performance analysis in a *modular* (particularly word-based) fashion.

A major factor making the word error rate difficult to interpret is the normalisation by $N_r$. While the edit distance (the numerator of the WER) itself has clear interpretation as an accumulated cost, normalising by the reference sequence length is problematic as the numerator is not bounded by $[0, N_r]$ due to the inclusion of insertions. This means that in practice the word error rate may exceed unity, bounded by $[0, I/N_r]$, or equivalently that the recognition rate may be negative. This property means that it is often difficult to interpret the meaning of the absolute value of the WER, or to make relative comparisons between two different rates. While we will not consider them further here, more principled approaches to normalised edit distances are presented in [2], and several alternate normalisations in a speech recognition context are mentioned in [3]. It is worth noting that the WCR above does not suffer from this property: the numerator is in fact bounded by the value in the denominator, yielding a true rate with clear interpretation (albeit one which does not consider insertion errors).

As well as being difficult to interpret, the WER does not allow thorough performance analysis according to varying word importance, as the string edit distance is not easily decomposed into measures per dictionary word. This limitation is due to the fact that the string edit distance considers three types of errors: insertions, deletions and substitutions. While insertion and deletion errors can clearly be associated with a single dictionary word, it is not clear to which word (either from the reference or automatic transcription) a substitution error should be attributed. As will be discussed later, the information retrieval framework advocated in this article instead considers only two classes of errors: inserted information (false alarms) and deleted information (false rejections).

Of course, such limitations may not be important in certain application or research contexts, and the WER may indeed provide a suitable evaluation measure in these cases. However, more general application-oriented performance analysis requires an alternative, or additional, evaluation framework.

## 2.2 Related Work

The recognised need for alternative ASR evaluation measures is certainly not novel to this article. A number of researchers have highlighted the above limitations of the WER and demonstrated that it is often not a good indicator of the usability of an end application. Perhaps one of the most striking examples can be found in [4], where it is shown that improvements in spoken language understanding can be obtained while observing a significant increase in the WER. Similarly, in spoken document retrieval applications, it has been repeatedly acknowledged that high word error rates do not necessarily lead to any significant degradation in retrieval performance, see e.g. [5, 6]. This has also been shown to be the case for spoken document clustering applications in [7].

Motivated by the evident limitations with the word error rate, several researchers have proposed alternative measures. In [3], the lack of a lower bound, and the consequent asymmetry with respect to insertions and deletions, were acknowledged as limitations of the WER and a new measure was proposed, termed *word information preserved* (WIP). This was derived as an approximate measure of mutual information between the reference and automatic transcriptions, and is given as $WIP =$

$H^2/[(H+S+D)(H+S+I)]$. While the derivations differ, it will become apparent that WIP has similarities to the approach presented in this article, as it can be seen as the square of the geometric mean of the precision and recall measures.

A work close to the current article is [8], in which a range of alternative ASR evaluation measures were proposed for a spoken document retrieval application. These measures were all based on the idea that a set of information-carrying words should be emphasised more in the word error rate measure so as to correlate better with eventual document retrieval performance (thus, application usability). The measures included: named entity word error rate, the stemmed stop-word-filtered word error rate, and the IR-weighted stemmed stop-word-filtered word error rate, amongst others. Significant correlation was verified between each of these measures and the eventual document retrieval performance. While the framework is somewhat different, these results provide experimental support for the approach to incorporate application dependent word importance weights, which we also advocate.

Another interesting article related to the current discussion is [9], in which the authors propose the adoption of the WER to measure information retrieval performance. In the following sections it will be apparent that we advocate the converse case: adopting recall and precision from information retrieval to assess speech recognition. The argument in [9] hinges on the point that the $F$ measure, commonly used in information retrieval to combine recall and precision into a single measure, effectively de-emphasises (by a factor of 2) the deletion and insertion errors with respect to substitution errors. In the present paper, we take the point of view that in fact each substitution produces two types of error, both a deletion (of the correct word) and an insertion (of the incorrect) word, and so in terms of its effect on the information content, it should be counted as such. In this context it is interesting to note that in [10] a weighted WER which explicitly de-emphasised deletion and insertion errors by a factor of 2 was proposed to avoid bias introduced in the dynamic programming alignment procedure (see Section 6.2).

# 3 Speech Recognition as a Case of Information Retrieval

In the preceding sections we have seen that the conventional word error (or recognition) rate is based on posing speech recognition evaluation as a string editing problem. In this article, we instead pose speech recognition evaluation as an *information retrieval* task, in which we treat each word occurrence as a *unit of information*, and in which the goal is for the *relevant* information present in the reference transcription to be *retrieved* in the automatic transcription. To avoid potential confusion, we note that this does not presume that the end application is necessarily itself a traditional information retrieval task.

## 3.1 Notation

Before proceeding, let us define the word vocabulary $V$ as a set of unique words, $V = \{v_i\}$, where $i$ is the word index ranging from one to the vocabulary size $|V|$ (where $|\cdot|$ denotes set cardinality). Let us also define the null word $\epsilon$, which is not a member of the vocabulary set. Assuming we have the word slot alignment (see Section 6.2 for some discussion), we have a sequence of $J$ word slots in both the reference $r = (r_{1:J}) = (r_1, \ldots, r_j, \ldots, r_J)$ and automatic $a = (a_{1:J})$ transcriptions, where $J$ equals the total number of words in the reference transcription plus the total insertions in the automatic transcription (equivalently, the total words in the automatic transcription plus deletions). In these sequences, insertion errors are represented as a null word slot in the reference transcription $r_j = \epsilon$, and deletions by a null word slot in the automatic transcription, $a_j = \epsilon$.

## 3.2 Evaluation Measures

In the following we show how ASR evaluation measures may be calculated per dictionary word, and per error type, using the concepts of recall and precision from information retrieval. In contrast to

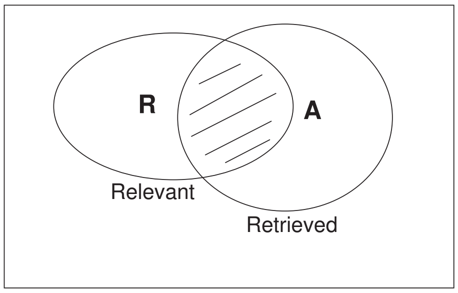



Figure 1: Venn diagram illustrating the concepts of recall and precision. Set $R$ is the relevant set from the reference transcription, and set $A$ is the retrieved set from the automatic transcription. Recall is the ratio between the cardinality of the intersection and the relevant set, while precision is the ratio of the cardinality of the intersection and the retrieved set.

the word error rate, each of the measures listed below is a true rate ranging in value between $[0, 1]$, and thus has clear interpretation in the context of a particular application.

#### 3.2.1 Word-based Recall and Precision

Consider that in an information retrieval context, the vocabulary $V$ defines the set of all possible queries (thus all queries consist of one vocabulary word). For a given vocabulary word, $v_i$, the set of relevant information units (word occurrences) is given as $R_i = \{j|r_j = v_i\}$, and the set of retrieved information units is given as $A_i = \{j|a_j = v_i\}$ (i.e., the non-null word slots from the reference and automatic transcriptions, respectively). The intersection of these two sets corresponds to the set of correctly recognised instances of the $i^{th}$ word, $R_i \cap A_i = \{j|r_j = a_j = v_i\}$. These are illustrated in Figure 1.

Measures most commonly used to evaluate information retrieval are the *recall* and *precision*, which are defined as [11]:

*Recall* is the fraction of the relevant information units (set $R_i$) which has been retrieved, i.e.,

$$\rho_i = \frac{|R_i \cap A_i|}{|R_i|}, \text{ and} \tag{4}$$

*Precision* is the fraction of the retrieved information units (set $A_i$) which is relevant, i.e.,

$$\pi_i = \frac{|R_i \cap A_i|}{|A_i|}. \tag{5}$$

These measures are only defined for words that appear in either the reference or automatic transcriptions, and both measures are defined as zero if they only appear in one of the transcriptions (either the relevant or retrieved sets).

#### 3.2.2 Global Recall and Precision

The above measures can be calculated for each word $v_i$, as above. To calculate the measures instead over the entire vocabulary, we may take either the *micro-average* or the *macro-average* [11]. The micro-average (denoted by subscript $\mu$) weighs each individual information unit (word occurrence) equally, as:

$$\rho_\mu = \frac{\sum_i |R_i \cap A_i|}{\sum_i |R_i|}, \text{ and} \tag{6}$$

$$\pi_\mu = \frac{\sum_i |R_i \cap A_i|}{\sum_i |A_i|}, \tag{7}$$

while the macro-average (denoted by subscript $M$) instead weighs each query (vocabulary word) equally:

$$\rho_M = \frac{1}{|V_r|} \sum_i \rho_i, \text{ and} \tag{8}$$

$$\pi_M = \frac{1}{|V_a|} \sum_i \pi_i, \tag{9}$$

where $V_r$ is the subset of words in $V$ that are present in the reference transcription, $V_a$ is the subset of words in $V$ occurring in the automatic transcription, and where in all cases the summation is only over terms $i$ where the corresponding word-based measures are defined.

#### 3.2.3 Combined Measures

While calculating both recall and precision measures offers the most flexible basis for performance analysis, it may sometimes be desirable to evaluate or optimise a system in terms of a single measure. The recall and precision measures can be combined in a single value in a number of ways. One common such measure is the $F$-measure, which is the harmonic mean of recall and precision [11]:

$$F_i = \frac{2 \cdot \pi_i \cdot \rho_i}{\pi_i + \rho_i}. \tag{10}$$

While common in information retrieval evaluations, the $F$-measure is by no means the only way of combining recall and precision into a single measure. A range of other measures are possible using different forms of averaging, and the choice of a particular combined measure should ultimately depend on the application.

The corresponding micro- or macro-averaged measures can of course be calculated in each case. Of all these measures, we note that the micro-average $F_\mu$ measure corresponds the closest to the word recognition rate, as it measures the performance over an entire word sequence, with each word occurrence being weighed equally.

### 3.3 Example

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Reference:** | The | cat | $\epsilon$ | sat | on | the | mat | at | the | door. |
| **Recognised:** | She | rat | the | sat | $\epsilon$ | the | mat | at | $\epsilon$ | door. |
| **Slot index $j$:** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Figure 2: Sample alignment of reference and recognised word sequences for calculation of evaluation measures. The symbol $\epsilon$ represents a null word (insertion or deletion error).

As a simple illustration of the proposed framework, consider the sentence shown in Figure 2. We can obtain a value of precision, recall, and any possible combined scores, for each word $i$ in the vocabulary. Taking, e.g., the word $v_i = the$, we have $R_i = \{1, 6, 9\}$ and $A_i = \{3, 6\}$, from which we can calculate: $\rho_i = 0.33$, $\pi_i = 0.5$, and $F_i = 0.4$. The interpretation of these measures is straightforward: a third of the occurrences of the word *the* in the reference transcription (relevant words) have been correctly recognised, while half of the occurrences of the word *the* in the automatic transcription (retrieved words) are correct.

Taking the micro-average across the entire vocabulary for the above sentence, we obtain: $\rho_\mu = \frac{5}{9} = 0.56$, $\pi_\mu = \frac{5}{8} = 0.63$, and $F_\mu = 0.59$. This shows us both that 56% of the word occurrences in the reference transcription are correctly recognised, and that 63% of the word occurrences in the automatic transcription are correct. Instead calculating the average measures over each vocabulary word, the corresponding macro-averages can be found to be: $\rho_M = 0.62$, $\pi_M = 0.64$, and $F_M = 0.63$. The word recognition rate in this case would be $WER = \frac{4}{9} = 0.44$.

# 4 Comparison with Word Error Rate

A major difference between the proposed information retrieval framework and the string edit framework of the WER is the way in which word substitution errors are handled. In our proposed framework there are fundamentally only two types of errors, insertions (false alarms) and deletions (false rejections): we view a substitution error as a construct describing the case when these co-occur[1]. In terms of information content, a substitution error represents both a loss of relevant information as well the retrieval of erroneous information, and thus is considered as both a deletion and an insertion error. While the “common sense” view in ASR considers that counting substitutions twice is unfair, it is evident that this should be the case if we consider properly the information content of the words in the context of an end application. Of course, it is feasible that the substitution of one particular word by another may be allowable for a given application as it incurs no cost in terms of system usability. Such cases can be catered for in the information retrieval framework by applying a text normalisation process (e.g. stemming, synonym-matching, homophone-matching) prior to calculating evaluation measures, as will be discussed briefly in Section 6.1.

With this in mind, relating the information retrieval framework to standard error types encountered in speech recognition, the word recognition rate is given by (from Eq. 2):

$$WRR = \frac{\sum_i (|R_i \cap A_i|) - I}{\sum_i |R_i|} = \rho_\mu - \frac{I}{N_r}, \tag{11}$$

which is a difficult quantity to interpret. This equation highlights the fact that the word recognition rate can be negative, and that interpreting values depends on the relative sizes of the relevant sets and the total number of insertions. We see here that the *WRR* can be considered as the micro-averaged recall penalised by including insertion errors in the numerator. In the information retrieval perspective presented here, there is no basis or clear interpretation for such a measure.

The *micro-averaged recall* can be written as

$$\rho_\mu = \frac{\sum_i |R_i \cap A_i|}{\sum_i |R_i|} = \frac{H}{N_r}, \tag{12}$$

which is equivalent to the word correct rate (WCR), and the *micro-averaged precision* can be expressed as

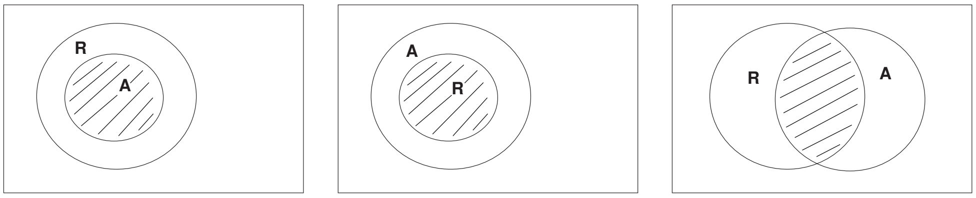
$$\pi_\mu = \frac{H}{N_a}. \tag{13}$$

From this we see that the WRR is essentially equivalent to the WCR (recall) penalised to also include insertion errors. A more consistent way of evaluating the rate of insertion errors is to instead define the corresponding precision measure, and use principled combinations, such as the $F$-measure, whenever a single measure is required.

For interest, we note here that the WIP (word information preserved) measure proposed in [3] can be interpreted as the product (squared geometric mean) of the micro-averaged recall and precision $WIP = \rho_\mu \cdot \pi_\mu$.

The Venn diagrams in Figure 3 illustrate the difficulties in interpreting word recognition rate. For simplicity, let us consider that in each of these cases all errors consist of word deletions or word insertions (i.e. no substitutions). Figure 3 (a) shows the case where all words in the automatic transcription are correct, but only half of the words in the reference were recognised (retrieved). It

[1] although note that the word alignment as commonly used in ASR evaluation does not necessarily guarantee temporal correspondence between aligned words. See Section 6.2 for discussion.




| (a) | (b) | (c) |
|---|---|---|
| $WRR = 0.5$ | $WRR = 0$ | $WRR = 0$ |
| $\pi = 1$, $\rho = 0.5$, $F = 0.67$ | $\pi = 0.5$, $\rho = 1$, $F = 0.67$ | $\pi = 0.5$, $\rho = 0.5$, $F = 0.5$ |

Figure 3: Venn diagram illustrating three cases of different relevant ($R$) and retrieved ($A$) sets, along with the measures that would result in each case.

| **Measure** | | **Recall** | **Precision** | **F** |
|---|---|---|---|---|
| Word-based | *trial* | 1.00 | 0.69 | 0.82 |
| | *university* | 0.59 | 1.00 | 0.74 |
| Average | micro | 0.63 | 0.69 | 0.67 |
| | macro | 0.54 | 0.58 | 0.56 |

Table 1: Evaluation measures calculated on CTS eval01 set, using the alignment provided by NIST scoring software. The corresponding WRR measure is 0.63 (WER of 37%).

can be seen that this case would lead to a word recognition rate of 0.5. Figure 3 (b) shows the case where all words in the reference transcription were recognised correctly, but this is only half of the total number of recognised (retrieved) words. It can be seen that this case would lead to a word recognition rate of 0. Figure 3 (c) shows the case where half the words in the reference transcription were recognised correctly, and this is only half of the total number of recognised (retrieved) words. This would again lead to a word recognition rate of 0. If we consider rather the precision and recall: in case (a) we would obtain $\pi = 1$, $\rho = 0.5$, and $F = 0.67$; in case (b) $\pi = 0.5$, $\rho = 1$, and $F = 0.67$; and in case (c) $\pi = 0.5$, $\rho = 0.5$ and $F = 0.5$. It is clear from this illustration that precision and recall are much simpler to interpret than word recognition rate with respect to an end application, and that as a combined single measure, values of the $F$- measure give a fair indication of the relative merits of the different systems.

Table 1 shows evaluation measures calculated on a test set (eval01) of the CTS (Switchboard) corpus. From the word-based measures, we can see that while the word *trial* has higher recall (more relevant words are retrieved) and a better combined $F$ measure, the word *university* has greater precision, meaning that when that word is observed in the automatic transcription, we can have a high confidence that it was actually spoken. While the ability to decompose the measure by word, or into recall and precision components is important in allowing us to better interpret results in this way, for many purposes it is also desirable to have a single combined measure. Table 1 shows that the averaged $F$ measures is a combined measure that can be used for similar purposes as the word error rate.

To illustrate the benefit in interpreting a measure that is a proper rate in the range $[0, 1]$, Figure 4 shows evaluation measures calculated on the Numbers 95 test corpus, varying as a function of the recogniser's word insertion log probability parameter. Increasing this parameter has the effect of making it 'easier' for the system to output a word in the automatic transcription. First looking at the recall and precision measures, we see the expected behaviour that the precision measure degrades with increased word insertion log probability, as more incorrect words are retrieved in the automatic transcription. The recall measure, on the other hand, remains relatively high, showing that the most of the relevant (correct) words from the reference transcription continue to be recognised in

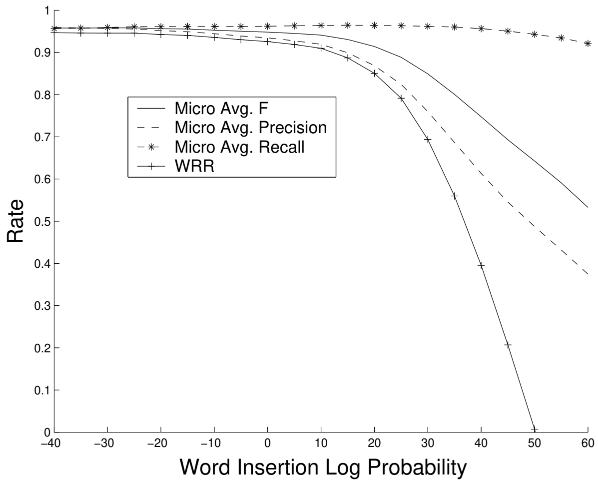



Figure 4: Evaluation measures calculated on the Numbers 95 test corpus, varying as a function of the recogniser's word insertion log probability parameter. Increasing this parameter favours having more words in the automatic transcription.

the automatic transcription. If we now look at the combined measures, we see that the $F$-measure always lies between the precision and recall curves, giving a meaningful combination of the two. On the other hand, the word recognition rate degrades rapidly with increasing insertion errors, dropping below both the recall and precision curves and quickly becoming negative.

# 5 Configuring Word Importance and Error Cost

The above framework gives a set of measures per vocabulary word, and per error type, allowing thorough performance analysis to be conducted in a given application context. In addition, when measures over the entire vocabulary, or over both error types, are required, these can be easy calculated by applying the form of averaging appropriate to the application. In the following, we briefly discuss the use of weighted averaging to incorporate application-dependent word importance and/or error costs in these combined measures.

Note that while we discuss a framework for incorporating such weights, it is beyond the scope of this work to determine what these weights should be for a particular application. The optimal balance between the relative importance of different words, or between types of error, may differ significantly between applications and is open to subjective determination. While simple weight selection schemes may be used, application performance is ultimately measured by effectiveness at satisfying users. Usability tests with human subjects are time-consuming and expensive, so are unrealistic in the day-to-day process of incremental ASR system development. However, whenever possible, some human-subject testing may be used to help determine the ideal weights. See [12] for an example of related work in this area.

## 5.1 Word Importance

The global precision and recall measures in Section 3.2.2 are averages taken over each vocabulary word. They reflect the case in which each word contributes equally to the system performance, weighed purely by the number of occurrences through the set cardinality. In the introduction, we proposed that a given task or application could be characterised (at least in terms of the requirements on its speech recognition component) through a set of importance weights for each vocabulary word.

Let us define a set of importance weights, say $w_i$ in the range $[0, 1]$, for each word $v_i$ in the vocabulary. We could then calculate weighted micro-averages as:

$$\rho_{\mu w} = \frac{\sum_i w_i |R_i \cap A_i|}{\sum_i w_i |R_i|}, \text{ and} \tag{14}$$

$$\pi_{\mu w} = \frac{\sum_i w_i |R_i \cap A_i|}{\sum_i w_i |A_i|}, \tag{15}$$

or the weighted macro-averages as:

$$\rho_{Mw} = \frac{1}{\sum_i w_i} \sum_i w_i \rho_i, \text{ and} \tag{16}$$

$$\pi_{Mw} = \frac{1}{\sum_i w_i} \sum_i w_i \pi_i, \tag{17}$$

where in each case the summation is only over terms $i$ where the corresponding word-based measures are defined. Combined measures, such as $F_{\mu w}$ and $F_{Mw}$, could naturally then be calculated from these measures. While not strictly necessary, if we constrain $\sum_i w_i = 1$, then these word importance weights can be understood as the prior belief we have on each word being relevant to our application. This gives us a framework in which a set of task-dependent importance weights can be introduced into our performance measure, while retaining an intuitive interpretation.

The word importance weights $w_i$ for a particular application may be determined in several ways, including: By the part of speech (e.g. all nouns have equal weight, others 0); Weighing common stop words with zero importance; Inverse document (word) frequency [11] over the corpus of reference transcriptions (where infrequent words are assumed to be more significant); and Inverse document frequency over some external corpus of relevance to the application domain. See [8] for some specific examples (using WER) in the context of a spoken document retrieval system.

### 5.2 Error Cost

The $F$-measure mentioned in Section 3.2.3 weighs recall and precision equally, and thus represents the case in which both insertion and deletion errors are equally costly to system usability. Similar combined measures exist which instead allow the relative costs of the error types to be configured. One such measure is the $E$ measure [11]:

$$E_i = 1 - \frac{(1 + b^2) \cdot \pi_i \cdot \rho_i}{(b^2 \cdot \pi_i) + \rho_i}, \tag{18}$$

in which $b$ gives a single parameter that indicates the relative importance of recall and precision. For $b = 1$, $E$ is the complement of the $F$ measure. Values of $b > 1$ indicate that the user is $b$ times relatively more interested in precision than recall, while values of $0 < b < 1$ indicate the converse.

The relative costs associated with the different error types (deletion or insertion) should relate to the real cost incurred by the errors. For instance, responding to a false alarm generated by a particular keyword detection system may cost the user significantly more effort (time, money) than the cost of ignoring a real alarm.

## 6 Further Considerations

### 6.1 Text Normalisation

In some end applications of speech recognition there may be no distinction between usability if, e.g., the word *governed* was recognised instead of the word *governing*. Current speech recognition evaluations consider this as much of an error as if the word *potato* was recognised.

The solution taken to this issue in information retrieval systems is to employ *stemming*, *thesaurus matching* and *homophone matching* techniques [11]. Stemming refers to the replacement of a word by its stem, which is the portion of a word that remains after removal of prefixes and suffixes: e.g, replacing both *governed* and *governing* by the stem *govern*. A range of standard stemming techniques exist [11]. Thesaurus matching involves replacing a word (or a set of words) by a synonym word (having similar meaning), and likewise homophone matching refers to the replacement of a word by another similar sounding word.

In general, this normalisation defines a mapping of the word $v_i$ to a word $\widehat{v}_k$, and the vocabulary $V = \{v_i\}$ to a normalised vocabulary $\widehat{V} = \{\widehat{v}_k\}$, with $|V| \geq |\widehat{V}|$. In our speech recognition evaluation context, this same mapping should be applied to normalise both the reference and automatic transcriptions before proceeding to calculate the evaluation measures (but after the actual recognition system has been run).

In summary, text normalisation techniques should be applied in the case that a substitution of one particular word (in the reference transcription) by another (in the automatic transcription) causes no penalty to the user of a given application. This is one important way in which the speech recognition evaluation measure can be made to better relate to end application usability.

### 6.2 Word Alignment

A key practical issue with word error rate calculation is finding the word alignment between reference and automatic transcriptions. This is commonly found using a dynamic programming algorithm to minimise the edit distance between the strings. An implicit assumption made in current speech recognition evaluation is therefore that the alignment found in this way is a good approximation (on average) of the actual temporal correspondence (in terms of both alignment and duration) between reference and recognised words. The issue of temporal correspondence is important: it is clear that if a system recognises the correct word at the wrong time, then we should not consider that the system functioned correctly even if we obtain good scores from our evaluation measures.

It is important to keep in mind that the minimum edit distance is used for two distinct purposes in current ASR evaluation: to find the word alignment via dynamic programming, and to score the aligned word sequences. The weights on insertions, deletions and substitutions (respectively $w_I$, $w_D$, $w_S$) used in finding the alignment are non-standard, although in general values such that $w_I = w_D$ and $w_S < w_I + w_D$ are used, see [3, 13]. Furthermore, these weights do not necessarily correspond to the costs used in the subsequent word error rate measure (which are generally all set to unity), meaning that, in general, the edit distance calculated for alignment differs from that used for evaluation. While it does give a reasonable approximation of the temporal alignment, the fact that the alignment procedure minimises an edit distance can introduce some bias into the WER measure, as observed in [10].

This is not to criticise current methods of finding the word alignment for evaluation, but merely to remind that this only gives an approximation to the true temporal alignment. This approximation may not be appropriate for certain applications or under certain conditions (such as high error rates), and thus can potentially lead to a distorted measure of system performance. Experiments demonstrating this, and investigating alternatives, can be found in [14, 15, 10].

## 7 Conclusion

Motivated by limitations of the commonly used word error rate, this paper has proposed a set of criteria necessary for effective, application-oriented evaluation measures of speech recognition. To progress toward the definition of such measures, the paper has defined speech recognition as an information retrieval task, in which each word is a unit of information, and in which the goal is for the automatic transcription to retrieve all the relevant information in the original speech signal. In such a framework, the familiar concepts of recall and precision offer measures with straightforward interpretation in a given application. Furthermore, this article has suggested that many end applications may be characterised by a set of importance weights associated with each vocabulary word, or costs associated with each error type, and has shown how these may easily be incorporated in the precision and recall measures, as well as any single-valued combinations of these. In this way, while the proposed measures are still direct and objective measures of speech recognition performance, they also yield straightforward interpretation, and allow thorough performance analysis to be done in a particular application context.

# 8 Acknowledgements

The authors wish to acknowledge the contribution of Alessandro Vinciarelli and David Grangier to discussions about the ideas presented in this article. This work was partly supported by the EU 6th FWP IST Integrated Project AMI (FP6-506811), the Swiss NCCR IM2, and the DARPA EARS programme.

# References

[1] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics - Doklady 10*, vol. 10, pp. 707–710, 1966.

[2] A. Marzal and E. Vidal, "Computation of normalized edit distance and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 926–932, September 1993.

[3] A. Morris, V. Maier, and P. Green, "From wer and wil to mer and wil: improved evaluation measures for connected speech recognition," in *Proceedings of International Conference on Spoken Language Processing*, 2004.

[4] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, December 2003.

[5] J. Garofolo, C. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the Eighth Text REtrieval Conference (TREC 8)*, November 1999.

[6] D. Grangier, A. Vinciarelli, and H. Bourlard, "Information retrieval on noisy text," IDIAP-COM 03-08, IDIAP, 2003.

[7] D. Grangier and A. Vinciarelli, "Noisy text clustering," IDIAP-RR 04-31, IDIAP, 2004.

[8] J. Garofolo, E. Voorhees, C. Auzanne, V. Stanford, and B. Lund, "1998 TREC-7 spoken document retrieval track overview and results," in *Proceedings of the Seventh Text REtrieval Conference (TREC 7)*, 1998.

[9] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proceedings of the DARPA Broadcast News Workshop*, February 1999.

[10] M. Hunt, "Figures of merit for assessing connected-word recognisers," *Speech Communication*, vol. 9, pp. 329–336, 1990.

[11] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 1999.

[12] C. Kamm, M. A. Walker, and D. Litman, "Evaluating spoken language systems," in *Proceedings of American Voice Input/Output Society, AVIOS*, 1999.

[13] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Engineering Department, 3.2 ed., 2002.

[14] W. Fisher and J. Fiscus, "Better alignment procedures for speech recognition evaluation," in *Proceedings of ICASSP '93*, vol. 2, pp. 59–62, 1993.

[15] M. Hunt, "Evaluating the performance of connected-word speech recognition systems," in *Proceedings of ICASSP-88*, vol. 1, pp. 457–460, 1988.