MEASURES OF LANGUAGE MODEL AND ACOUSTIC MODEL 
INFORMATION IN PROBABILISTIC SPEECH RECOGNITION

Marco Ferretti, Giulio Maltese, Stefano Scarci

IBM Rome Scientific Center
via Giorgione 159, 00147 ROMA (Italy)

ABSTRACT

To predict the performance of a probabilistic speech recognizer it is 
often desirable to estimate the contribution of the language model 
and that of the acoustic model. We describe an approach to this 
problem which tries to take into account the interaction between 
the two sources of information. Some results are presented concerning 
the 20000-word vocabulary, real-time IBM recognizer of the Italian 
language.

INTRODUCTION

The performance of a speech recognizer is usually measured by the 
experimental accuracy displayed on a sample of sentences uttered by 
a few speakers. It is often desirable to predict the behavior of the 
speech decoder on arbitrary (and possibly large) texts, without having 
them available in recorded form. It is also sometimes interesting to 
evaluate the individual contribution of the different components of 
the decoder. Let \( A \) be the acoustic information used by the speech 
decoder and \( W \) a sequence of words. In probabilistic speech 
recognition, the acoustic model (which estimates \( P(W|I) \)) and the 
language model (which estimates \( P(I) \)) can be developed separately. 
It is therefore desirable to measure separately their performance.

The effectiveness of the language model is usually expressed by 
means of its predictive power, as measured by perplexity [1]. 
Unfortunately, because this measure does not take into any account 
the relationship to the acoustic model, it cannot be assumed that a 
decrease in perplexity necessarily results in more accurate speech 
decoding. Intuitively, if two words are easily distinguished by their 
pronunciation, the information provided by the context to 
distinguish them is of much less relevance to the accuracy of the 
speech decoder than for two acoustically confusable words. In [2] it 
is presented an example of a language model based on parts of 
speech which has higher perplexity of a trigram model, but achieves 
more accurate recognition on a test text uttered by a speaker. Peter 
De Souza proposed a measure, called aperplexity, described in [3], 
which attempts to take into account acoustic/language models 
interaction. Aperplexity is the average perplexity of the language 
model when its choice is limited to a subset of the vocabulary made 
out of the words which are acoustically most similar to the correct word.

As reported in [4], the perplexity of a trigram language model, when 
it's choice is limited to a random subset of \( m \) words of the vocabulary 
(including the correct word), was measured for varying \( m \), for a 
6000-word Italian vocabulary. The same experiment, performed on 
subsets selected according to acoustic similarity to the right word, 
showed no significant differences in the behavior of the aperplexity as 
a function of \( m \). This result suggests that, if aperplexity is assumed 
to describe significantly enough the interaction between the language 
and the acoustic model, in that specific case the interaction was 
essentially irrelevant and perplexity was a measure as good as 
aperplexity of the contribution of the language model to the 
performance of the speech decoder.

In [5] we proposed a new measure, called Speech Decoder Entropy 
(SDE), of joint acoustic-context information. It is more expensive to 
compute than perplexity and aperplexity, but we believe it provides a 
better understanding of the interaction of the acoustic and language 
models, and a better prediction of the contribution of a model to the 
performance of the speech decoder.

The paper is organized as follows. In the next section we recall 
definitions for SDE and some related quantities. The following 
section contains a brief description of the IBM 20000-word 
recognizer for the Italian language. In the following section, we 
report some recent experimental results about acoustic/language 
model information measures. In the last section we discuss future developments.

SPEECH DECODER ENTROPY

We shall use the following notation:

- \( W \) word to be decoded
- \( C \) linguistic context
- \( A \) acoustic data for \( W \).

We shall consider the amount of information on the word to be 
decoded supplied separately by the context and the acoustic data, and 
the amount of information supplied jointly:

\[
I(W;C) = H(W) - H(W|C) \\
I(W;A) = H(W) - H(W|A) \\
I(W;A,C) = H(W) - H(W|A,C).
\]
The figure most significantly related to the performance of the decoder is, presumably, $H(W|A,C)$, which we call Speech Decoder Entropy.

The number:
\[ \Delta = I(W;A) + I(W;C) - I(W;A,C) \]

expresses the loss of mutual information when using jointly the context and the acoustic, with respect to using them separately. By observing that
\[ I(W;A) - I(W;A,C) = - I(W;C|A) \]

we obtain another expression for $\Delta$:
\[ \Delta = I(W;C) - I(W;C|A) \]

which allows to interpret $\Delta$ as the loss of mutual information between the context and the word, when acoustic information on the word is also given. Being the expression of $\Delta$ symmetrical in $A$ and $C$, another interpretation can be given by exchanging acoustic and context in the previous sentence. Intuitively, the smaller is $\Delta$, the more "orthogonal" are the information provided by the two models.

Still another interesting expression for $\Delta$ can be obtained if we assume that the context $C$ influences the utterance $A$ only through the word $W$, i.e.
\[ P(A|W,C) = P(A|W). \quad (1) \]

This assumption is especially reasonable for speech made out of discrete word utterances. As derived in [5], it is found
\[ \Delta = I(A;C). \]

This shows that an increase of $I(W;C)$ (or $I(W;A)$), due to a variation of the language (or acoustic) model, does not necessarily result in an equal improvement of the overall information $I(W;A,C)$, because it might rather increase the information between context and acoustics, $I(A;C)$, which is useless for the purpose of decoding $W$.

**THE IBM SPEECH RECOGNIZER FOR ITALIAN**

While the estimation techniques we propose are general, and can be applied to any statistical speech recognizer, some points will be referred to the approach followed in the IBM recognizers [6] and the experimental results will regard the speech recognizer for the Italian language [4][7].

Acoustic probabilities are computed by means of Hidden Markov Models. Both phonetic (based on an enlarged set of 56 Italian phones) and fenonic [8] models are used. The speaker trains the parameters by reading a 15-minute long phonetically-balanced training text. Short pauses have to be left between words. The acoustic match is performed in two stages, the first of which (called fast match) selects by means of a fast, approximate computation, a subset of the vocabulary made out of the acoustically most likely words, while the second one (called detailed match) computes more accurate acoustic probabilities for the selected words.

Language probabilities are estimated by a tri-gram language model, based on counts collected from large corpora. Corpora are currently built from Italian text made out of news on economy and finance. In order to deal with unseen events (a problem arising from the finite size of the training corpora) frequencies computed from counts must be smoothed. Two techniques have been used:

- **Back-off:** when no tri-gram count is available for the context, the model backs off to a bigram-based prediction, and if a bigram count is not available either, to an unigram-based prediction. The probability of unseen events is computed according to Turing's formula [9].
- **Deleted interpolation:** the probability is based on a linear combination of trigram, bigram and unigram predictions. Weights are computed by maximizing the likelihood of held-out data [10].

**EXPERIMENTAL RESULTS**

Let $w_1, \ldots, w_t$ be a text $T$ of $t$ words; $c_t$ the linguistic context in which $w_t$ occurs (for a trigram language model, for example, $c_t = w_{t-2}, w_{t-1}$); $a_{w_t}$ the acoustic information relative to the utterance of word $w_t$ by a speaker who has recorded text $T$. Assuming proper ergodic behavior, the measures of interest can be estimated as follows.

\[ \tilde{H}(W) = - \frac{1}{t} \sum_{k=1}^{t} \log \tilde{P}_u(w_k) \]  

(2)

where $\tilde{P}_u$ is the unigram probability.

\[ \tilde{H}(W|C) = - \frac{1}{t} \sum_{k=1}^{t} \log \tilde{P}_l(w_k|c_k) \]  

(3)

where $\tilde{P}_l$ is the language model probability (a trigram model in our case).

\[ \tilde{H}(W|A) = - \frac{1}{t} \sum_{k=1}^{t} \log \left( \sum_{w_k \sim \text{int}(a_{w_k})} S_u(w_k, a_{w_k}) \tilde{P}_u(w_k) \right) \]  

(4)

where $S_u(w, a)$ is the acoustic model score, which is proportional, for a given utterance $a$, to $P(a|w)$. Acoustic scores are raised to an exponent of 0.25 to balance the underestimation of the probability by the HMMs [5]. The sum at the denominator is extended to the acoustically most likely words, as found by the acoustic fast match. This reduces computation time, without affecting significantly accuracy, because acoustic scores of words not part of the fast match list sum up to a negligible amount.
In the same way:

\[
\widetilde{H}(W|A,C) = \frac{1}{l} \sum_{i=1}^{l} \log \frac{S_a(w_i, a_n) \tilde{P}_r(w_i | c)}{\sum_{w \in \text{vocabulary}_{a_n}} S_a(w, a_n) \tilde{P}_r(w | c)}
\]

(5)

Aperture can be computed using the same expression, where acoustic model scores are assumed uniform.

By making use of the assumption (1), similar formulas can be used for obtained estimates on arbitrary text, not available in recorded form (we use a database of utterances and acoustic scores covering the whole 20000-word vocabulary for this purpose).

Six language models were used in the measurements. They differ for:

• the smoothing technique;
• the size of the corpus from which the trigram and bigram statistics were taken;
• the threshold on the number of occurrences (only bigrams and trigrams which occur more than a minimum number of times in the corpus are included in the statistics).

The models are described in Table 1. Corpora sizes are in millions of words; the number of bigrams and trigrams is in millions; Bx models are smoothed by backing-off, D0 is smoothed by deleted-interpolation.

Table 1. Language models.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Threshold</th>
<th>Corpus Size</th>
<th>Different n-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0</td>
<td>1 2G 21</td>
<td>1 2G 21</td>
<td>0.8 1.9</td>
</tr>
<tr>
<td>D0</td>
<td>0 1 21</td>
<td>0 1 21</td>
<td>1.7 1.9</td>
</tr>
<tr>
<td>B1</td>
<td>1 1 40</td>
<td>1 1 40</td>
<td>1.2 3.1</td>
</tr>
<tr>
<td>B2</td>
<td>1 1 107</td>
<td>1 1 107</td>
<td>2.0 3.1</td>
</tr>
<tr>
<td>B3</td>
<td>1 2 107</td>
<td>1 2 107</td>
<td>2.0 4.4</td>
</tr>
<tr>
<td>B4</td>
<td>0 1 107</td>
<td>0 1 107</td>
<td>4.0 2.1</td>
</tr>
</tbody>
</table>

Smaller corpora are subsets of larger ones.

We report some measures for a test T for which acoustic data by some speaker are available. It is the standard text we use for test decodings. It consists of 61 sentences, amounting to 1024 words. For this text \(H(W|A)=9.705\) bits (in the following all entropies are expressed in bits).

We used five recordings of the text by five different speakers. The acoustic probabilities were computed by the detailed match, \(H(W|A)\) was estimated by (4), where, for each word \(w_i\), the logarithm of the probability was averaged on the five speakers. We obtained \(H(W|A)=0.400\). Table 2 reports values of SDE, LME (i.e. \(H(W|C)\)), AP (aperture), \(A\), NERR (the cumulative number of errors made by the recognizer when decoding text T uttered by a disjoint set of seven speakers), for the six language models.

Table 2. Relevant measures for the six language models.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>LME</th>
<th>AP</th>
<th>SDE</th>
<th>(A)</th>
<th>NERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0</td>
<td>6.832</td>
<td>1.932</td>
<td>0.285</td>
<td>2.758</td>
<td>402</td>
</tr>
<tr>
<td>D0</td>
<td>6.840</td>
<td>1.917</td>
<td>0.281</td>
<td>2.746</td>
<td>402</td>
</tr>
<tr>
<td>B1</td>
<td>6.741</td>
<td>1.859</td>
<td>0.276</td>
<td>2.840</td>
<td>380</td>
</tr>
<tr>
<td>B2</td>
<td>6.671</td>
<td>1.812</td>
<td>0.270</td>
<td>2.904</td>
<td>371</td>
</tr>
<tr>
<td>B3</td>
<td>6.619</td>
<td>1.783</td>
<td>0.267</td>
<td>2.953</td>
<td>364</td>
</tr>
<tr>
<td>B4</td>
<td>6.712</td>
<td>1.822</td>
<td>0.274</td>
<td>2.867</td>
<td>372</td>
</tr>
</tbody>
</table>

Since all the models share essentially the same structure, we cannot expect large or unpredictable differences in any of the measures. For the backing-off models, the larger is the corpus, the higher is the recognition rate, and this is reflected by all the measures (LME, AP, SDE). It is of some interest to compare B0 to D0: they were built from the same data, but with different thresholds and smoothing strategies. While the predictive power of B0, as measured by LME, is higher, its interaction with the acoustic model is not as good as that of D0 (as shown by the higher \(A\)), and they result in the same number of errors. The finding that the backing-off strategy achieves slightly better perplexity is consistent with what reported in [5]; our data show that, at least in this case, this does not imply better recognition accuracy. Notice that SDE, \(A\), and AP are computed from acoustic data disjoint from that used to compute NERR.

CONCLUSIONS

We have described some techniques for the estimation of the performance of the language model in a probabilistic speech recognizer, which take into account the interaction with the acoustic model, first proposed in [5]. We have presented some new experimental results regarding the comparison of trigram models built using different smoothing strategies. While in the case considered the value of \(A\) gave account of the fact that the model with better perplexity did not achieve higher recognition accuracy, the differences involved are too small to draw definitive conclusions.

We believe that measurements on language models with very different structures have to be performed in order to assess whether the proposed measures can give an improvement to the estimation of the contribution of the language model to the speech decoder, large enough to justify their increased computational cost with respect to perplexity.

EUROSPEECH '89, Paris, France, September 1989

2475
REFERENCES


