INVESTIGATION OF LANGUAGE STRUCTURE BY MEANS OF LANGUAGE MODELS INCORPORATING BREATHING AND ARTICULATORY NOISE

Karl Weilhammer* and Florian Schiel†
**Department of Phonetics and Speech Communication, University of Munich*
†Bavarian Archive for Speech Signals, University of Munich

ABSTRACT

In our experiment we used a bigram language model and a standard speech recogniser to test if linguistic information is related to the position of silence, articulatory noise, background noise, laughing and breathing in spontaneous speech.

We observed that for silence and articulatory noise the acoustic modelling is more important than linguistic information represented in the bigrams of a language model. Breathing carries useful information that can be described in a language model, because including it into the language model improves test set perplexity and recognition accuracy.

This means that precisely defined noise items add some linguistic knowledge to the language model and contribute to a better performance of an automatic speech recogniser.

1. INTRODUCTION

Language models using bigram statistics are standard tools in automatic speech recognition (ASR). Usually only the bigrams of orthographic units are considered. In recent years other authors have drawn their attention to modelling hesitations and disfluencies in spontaneous speech additional to lexical words [7][8].

In this article we will discuss the influence of breathing, articulatory noise, laughing, background noise and silence on the communication process. Our hypothesis was that the perplexity of a language model containing an item which carries linguistic information should decrease compared to a model containing only words, thus possibly improving speech recognition. To evaluate the influence of the acoustical modelling vs. the language model we conducted several experiments on a standard HMM speech recognition system.

Our investigation is based on the VERBMOBIL 1 Corpus consisting of spontaneous speech dialogues uttered by German native speakers.

The following section describes the data and the ASR system used in our experiments, while the third and fourth section discuss the usage of silence and other noise categories respectively.

2. EXPERIMENTAL SETUP

2.1 The Verbmobil 1 Corpus

During the first part of the VERBMOBIL project 789 spontaneous speech dialogues have been recorded and transliterated [1].

In all the recordings two German native speakers were asked to fix a date for a business meeting [2] by using previously prepared diaries. The scenario was modified slightly in different ways. In the simplest case they only had to fix up to three dates on a business trip or for a short meeting. More complicated tasks included arranging a flight or a train to their appointment or were even complete travel planning. The dialogues were recorded in Kiel, Bonn, Karlsruhe and Munich, which means that the vocabulary of each site is influenced by different regional variants of standard German.

Besides orthographic items breathing, speech pauses, four different classes of hesitations, six classes of articulatory noise and four classes of background noise have been carefully annotated by human transcribers [3]. For our investigation we merged some of these classes and ended up with the set of items displayed in Table 1.

breathing
articulatory noise (swallow, cough, smack)
laughing
background noise (knock, buzz, rustle)
silence

Table 1: The set of noise items used in our investigation

For this investigation we used 729 dialogues containing 329897 uttered words (inclusive hesitations) and 62513 noise items to train the speech recogniser and to build the language models. The test corpus consisted out of 64 dialogues (15043 words and 3221 noise items).

2.2 The Recogniser

The recogniser uses the Hidden Markov Model Tool Kit (HTK), which is commercially available by Entropic [6]. The acoustical signal is transformed into a 39 dimensional Vector, that consists of logarithm of Energy, 12 mel-cepstral coefficients, their delta and delta-delta coefficients. The vectors are calculated every 10 msec.

The HMM modelling is based on a set of 48 phonemes including regular phonemes and the noise items listed in Table 1. We trained 1075 triphone models for phoneme combinations which occurred more often then 150 times in the training corpus. The remaining triphones were modelled by simple monophones. The HMMs have three to four emitting nodes with single gaussians modelled by full covariance matrices.

We used a back-of bigram language model, that takes the

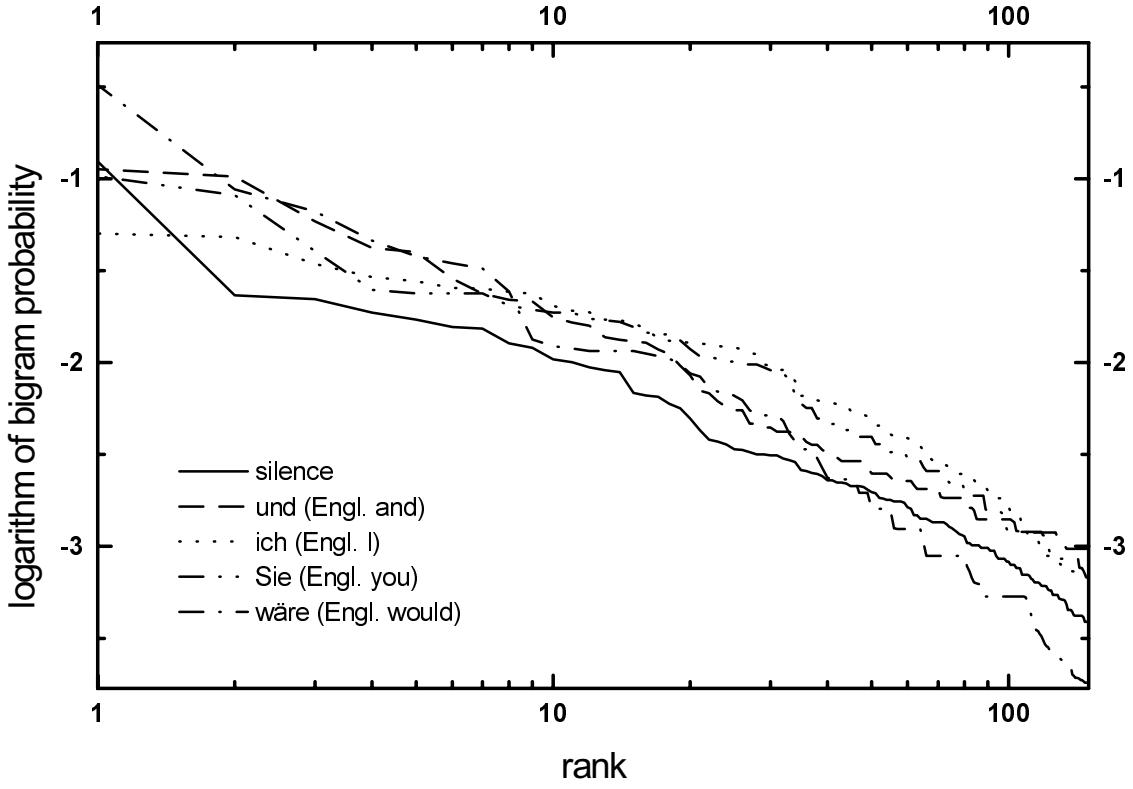

Figure 1: Bigram distribution of the silence item and some frequent words in the corpus.

bigram probability, if at least one bigram has been observed in the training data, otherwise it calculates the transition probability from the unigram count. The vocabulary size was 1263 items, based on the wordlist of the transcription of the test data. Only for these words bigrams have been included in the language model. Silence and noise items are not evaluated in the word accuracy of the recognition.

3. THE SILENCE MODEL

The quality of the silence model has a significant influence on the performance of the recogniser. We compared three different ways of silence modelling.

In the first system every word is followed by a one-state silence model that may be totally bypassed (t-model). The second system has a more elaborate silence model after each word. It consists out of a t-model that can optionally be followed by up to three instances of a three-state silence HMM that was trained to labelled silence portions of the training corpus. In both systems there is no entry for silence in the language model.

In the third system every word is followed by a silence t-model and the above described three-state silence HMM was added to the language model. The word recognition accuracies obtained by these three systems are given in Table 2 (Accuracy is defined as (no. of words – deletions – insertions – replacements) / no. of words).

SYSTEM	ACCURACY
simple silence t-model after each word	56.64%
complicated silence model after each word	65.00%
simple silence t-model after each word and silence in the language model	64.78%

Table 2: Recognition accuracies for different silence modelling.

The performance of the second and the third system is almost identical, but far better than that of the first one. Another experiment with the complex silence model after each word and silence in the Language model reached 65.01% accuracy. That means the language model does not contain any linguistic information related to the position of speech pauses that would improve the recognition accuracy.

According to Table 3 the perplexity per transition increases to a great extend when the silence notation is included into test set and language model. This is due to the distribution of bigram probabilities P(word|silence). Silence is with 32129 hits the most frequent item in the training corpus. Nearly half of the transitions

to the words in the test corpus have been observed in the training data. Therefore the entropy of the bigram distribution of silence is amongst the lowest 20% of all items of the lexicon.

Figure 1 displays the bigram distributions of silence and of some frequent words of the corpus. For the 50 most likely transitions the line for silence is extremely low compared to the others. This is the reason for the increase in perplexity caused by the silence model.

The bigram distribution of silence in the language model is not even approximately uniform. In system 2 it is equally likely that each word is followed by the silence t-model, one three-state silence HMM, or even two three-state HMMs. All these findings suggests that the gain in recognition accuracy must be due to good acoustic modelling.

4. THE NOISE CATEGORIES

For all the following investigations system two with the complex silence model and only words and hesitations in the language model will be regarded as the baseline system. The complex model represents the best silence modelling and was used in all following tests.

LANGUAGE MODEL	PERPLEXITY
words only	71.05
words, articulatory noise	71.25
words, breathing	68.40
words, laughing	71.15
words, background noise	73.85
words, silence	89.03
words, articulatory noise, breathing	68.37
words, articulatory noise, breathing, laughing	68.44
words, articulatory noise, breathing, laughing, background noise	70.57

Table 3: Perplexity per transition, calculated for different language models. The perplexity was derived from the same basis test set to which the relevant noise annotation was added. The language models were derived from the training set with the specified annotation added.

4.1 Background Noise

Some of the background noise is caused by the speakers, when they knock on the table, move their seats or rustle with paper. Some subjects were observed to knock on the table in the rhythm of their speech.

The test set perplexity increases when background noise is included into the language model. And the recognition accuracy does not substantially change when background noise is supported in the language model. It occurs 7413 times in the training corpus and 492 times in the test set and is therefore the fifth frequent item in the data base.

Since we discarded the noise markers that superimposed words, background noise was only observed within silence and is therefore nothing more than an inefficient silence model.

4.2 Laughing

Laughing is a very rare item, it occurs 239 times in the training corpus and only twice in the test set, therefore it does not alter recognition accuracy and test set perplexity substantially.

4.3 Articulatory Noise

Articulatory noise is caused by swallowing, coughing or smacking. It can be found amongst the ten most frequent items and occurs 5889 times in the training corpus and 328 times in the test set. Included into the language model it causes an insignificant increase in the test set perplexity and a gain of 0.62% in recognition accuracy. This suggests that articulatory noise carries not much linguistic information related to its position in spontaneous speech and the increase in recognition accuracy comes largely from the acoustic modelling.

4.4 Breathing

There are many publications in phonetic journals that deal with breathing during speech. [4] and [5] report that inspirations are largely taken at sentence boundaries or other positions appropriate to the grammatical structure of spontaneous speech.

Breathing is with 16843 hits in the training corpus and 771 hits in the test corpus the second frequent item in both sets. It causes an insignificant decrease in the test set perplexity and a gain of 0.52% in recognition accuracy. Therefore breathing contributes indeed a bit of linguistic information related to its position in spontaneous speech to the language model.

LANGUAGE MODEL	ACCURACY
baseline model, words only	65.00
words, background noise	65.23
words, laughing	65.22
words, breathing	65.52
words, articulatory noise	65.62
words, articulatory noise, breathing	65.73
words, articulatory noise, breathing, laughing	65.72
words, articulatory noise, breathing, laughing, background noise	65.71
words, articulatory noise, breathing, laughing, background noise, silence	65.76

Table 4: Recognition accuracies for different language models. The recognition for all systems was performed on the same test corpus and the recogniser used the complex silence model

4.5 Combinations

The language model which contains breathing and articulatory noise has the lowest test set perplexity in our experiment and increases the recognition accuracy by 0.73%

Including all noise items and silence in the language model improves the recognition accuracy to 65.76% compared to the baseline model with 65.00%. This difference is significant at the 0.1 level.

A recognition system with the simple silence t-model at the end of each word and all noise items plus silence in the language model reaches 65.69% recognition accuracy.

5. CONCLUSION

Our hypothesis that linguistic information related to the position of non-verbal items in spontaneous speech is reflected in a bigram language model was confirmed only for breathing. This result encourages some effort to improve the acoustical model of breathing.

Articulatory noise and silence did not improve the perplexity of the language model on the test set. One possible explanation could be that both items do not carry linguistic information; the other is that the linguistic information carried by these items cannot be represented in a bigram statistic.

For instance, articulatory noise is a mixture of swallowing, coughing or smacking. If these sub-categories each hold different types of linguistic information, the distribution is blurred and therefore spoils the test set perplexity. The same effect could hold for silence, which cannot easily be split off into sub-categories. Therefore a complex silence model after each word is more effective than including silence into the language model.

To get more reliable results a bigger corpus should be used for training and testing.

ACKNOWLEDGMENTS

We want to thank all the students that carefully transliterated the recorded dialogues. The results of our work rely on the accurateness of their transliterations.

REFERENCES

[1] Wahlster, W. 1997. Verbmobil: Übersetzung von Verhandlungsdialogen. *Verbmobil report-01-93*, available via Internet: http://www.dfki.de/cgi-bin7verbmobil/htbin/doc-access.cgi

[2] Kohler, K., Lex, G., Pätzold, M., Scheffers, M., Simpson, A., Thon, W. 1994. Handbuch zur Datenerhebung und Transliteration in TP14 von Verbmobil- 3-0. *Verbmobil Techdok-11-94*.

[3] Burger, S. 1997. Transliteration spontansprachlicher Daten, Lexikon der Transliterationskonventionen Verbmobil II. *Verbmobil Techdok-56-97*, available via Internet:
http://www.phonetik.uni-muenchen.de/VMTechDocs.html

[4] Winkworth, A. L., Davis, P. J. 1995. Breathing Patterns During Spontaneous Speech. *Journal of Speech and Hearing Research*, Vol. 38, pp. 124-144.

[5] Henderson, A., Goldman-Eisler, F., Skarbek, A. 1965. Temporal Patterns of cognitive Activity and Breath control in speech. *Language and Speech*, 8, pp. 336-242.

[6] Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P. 1997. The HTK Book (Version 2.1), Cambridge University

[7] Shriberg, E., Stolcke, A. 1996. Word Predictability after Hesitations: A Corpus-Based Study. *Proc. Intl. Conf. on Spoken Language Processing Philadelphia* pp. 1886-1871.

[8] Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauche, M., Tür, G., Lu, Y. 1998. *Proc. Intl. Conf. on Spoken Language Processing* Sydney, Vol 5, pp. 2247-2250.