Statistical Language Models for Large Vocabulary Spontaneous Speech Recognition in Dutch

Jacques Duchateau, Dong Hoon Van Uytsel, Hugo Van hamme, Patrick Wambacq

Katholieke Universiteit Leuven - ESAT
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium

E-mail: Jacques.Duchateau@esat.kuleuven.ac.be

10.21437/Interspeech.2005-22

Abstract

In state-of-the-art large vocabulary automatic recognition systems, a large statistical language model is used, typically an N-gram. However in order to estimate this model, a large database of sentences or texts in the same style as the recognition task is needed. For spontaneous speech one doesn't dispose of such database since it should consist of accurate thus expensive orthographic transcriptions of spoken audio.

This paper investigates how readily available large news paper corpora can be used to improve language models for spontaneous speech recognition although both language styles differ considerably. A technique is proposed that does a perplexity based automatic selection of appropriate news paper articles and that subsequently uses these texts in the language model estimation. Recognition experiments on spontaneous broadcast speech in Dutch showed significant improvements using this technique.

1. Introduction

Spontaneous speech recognition is currently a hot topic in speech research. This is not surprising as a wide range of practical applications based on automatic recognition of spontaneous speech become feasible: voice operated telephone services, automatic closed captioning for TV programmes, automatic transcription of meetings, etc. Yet, the recognition accuracy of freely spoken language is quite poor when compared to that of dictated speech: while the word error rates (WER) for large vocabulary speaker-independent dictation and broadcast news transcription are of the order of 5% [1] and 15% [2, 3] respectively, the WER for meeting and telephone conversation transcription [4] usually amounts to 40% or more.

Apart from the problems at the acoustical level that are due to a typically low audio quality and a sloppy pronunciation, the main reason for this discrepancy is the worse prediction capability of the language model involved in the recognition. This is mainly due to the occurrence of disfluencies in casual speech and to the lack of a sufficient amount of stylistically matching training data to estimate spontaneous language models. Several specific solutions to the disfluency problem were proposed in the literature, for instance our own previous work on the topic [5]. In this paper we try to overcome the second problem, the rather small training databases for spontaneous speech.

Stylistically correct training material for spontaneous speech consists of accurate (e.g. including the disfluencies) written transcripts of casual language. These manual transcripts are expensive and therefore rather scarce, typically not more than a few million words. On the other hand, typical large vocabulary statistical language models for dictation rely on vast amounts of training material [6], typically hundreds (and these days thousands) of millions of words from news papers. Unfortunately there is an obvious stylistical difference between this data and spontaneous speech. When building a spontaneous language model, using only the spontaneous speech transcripts leads to inaccurate statistics while using only the news paper material leads to incorrect statistics.

A solution to this problem is to use both sources at the same time. However simply putting together the statistics is not good as the statistics from the spontaneous speech transcripts may submerge in those from the written text material. Or at least the statistics of the added newspaper data may worsen the statistics from the spontaneous speech transcripts. Therefore in this paper a technique is proposed that selects the news paper material that resembles most to spontaneous speech, and that uses the selected articles to augment the spontaneous speech transcripts as language model training material.

The paper is organised as follows. First, we discuss how to select the *spontaneous* articles from the news papers. Next a description is given of the databases and the baseline recogniser involved in the experiments. Finally the experimental results with the proposed language modelling technique are given and discussed.

This research was supported by IWT projects ATraNoS (stww/000151) and SPACE (sbo/040102).

2. Measures for style match

Automatic selection of relevant documents from a large document set is a well-studied engineering topic that leads to well-known applications such as web search, document classification, and information retrieval. However, most of these select documents for a certain *topic*, rather than a certain *style*. For instance, information retrieval methods commonly rely on the TF×IDF feature [7]. The IDF weight (inverse document frequency) cancels the influence of frequent words, which we expect to indicate style (e.g. *you* and *me* in conversations).

On the linguistics side, the quantitative description of (writing) style is called stylometry. Traditionally, the focus was on the authorship of a text. More recently, the field was broadened [8] to the description of style and genre differences in English using factor analysis applied to computable linguistic features of texts. Automatic style detection based on this work is possible, but requires a lot of natural language processing (tagging, parsing, anaphora resolution, etc.) in order to compute the proposed features.

In this paper, a simpler approach is investigated. We used perplexity, or equivalently empirical cross-entropy with respect to an n-gram language model m as a measure of the distance between m and the document d:

$$\log \text{PPL}_d = -\sum_{w_1,..,w_n} f(w_1,..,w_n;d) \log p_m(w_n|w_1,..,w_{n-1}),$$

where the sum ranges over all n-grams, $f(x;d)$ gives the relative observation frequency of x in document d, and $p_m(w|h)$ returns the probability that context h is followed by word w according to the language model m. The influence of infrequent n-grams, which are likely to convey topic rather than style, can be reduced by simply omitting their corresponding terms from the sum in the above equation. The PPL measure is not biased by the length of d because the frequencies $f(x;d)$ are normalised.

This approach is actually the reverse of the probabilistic information retrieval method proposed in [9], which builds a language model m_d for each document d and estimates the probability of the query according to m_d.

3. Databases

This section shortly describes the databases involved in the language (and acoustical) modelling in the experiments in this paper.

3.1. News paper corpora

Two news paper corpora were used in the experiments. The *De Standaard* corpus contains about 35M words from a Flemish news paper of the same name, articles from years 1994, 1995 and 1996. The *Mediargus* corpus includes data from several Flemish daily and weekly papers in years 1999 and 2000, totalling almost 400M words in over 1M articles.

3.2. The CGN corpus

For the experiments, the CGN corpus was used (Corpus Gesproken Nederlands, Spoken Dutch Corpus [1]). In CGN, 3 types of spontaneous speech can be found: face to face dialogues, dialogues over telephone, and broadcast data. In addition, a difference can be made between the data collected in Flanders (one third, called *Flemish*) and the data collected in The Netherlands (two thirds, called *Dutch*).

From the Flemish CGN broadcast data (radio and television programmes), a 5k word test set was selected. This data was of course always excluded both from acoustic model and from language model training data.

4. Baseline recognition system

4.1. System overview

The large vocabulary speech recognition system used for the experiments was developed at ESAT [10, 11].

The acoustic model was estimated on 44h of Flemish spontaneous speech from CGN. Phonetic decision tree based context dependent models resulted in about 3500 tied states, modeled as a mixture of on average 100 tied gaussian distributions out of a total set of 32k different gaussians. No speaker adaptation was applied.

The lexicon consists of the 40k most frequent words in the *De Standaard* corpus for which a phonetic transcription can be found in Fonilex pronunciation database [2]. Five interjections, described phonetically (without word-dependent acoustic modelling), were added to the lexicon. Given this lexicon, the 5k word test set is characterised by a 3.5% OOV rate and a 6.2% interjection rate.

In the experiments, different options for the language model training data were investigated. Each time, a Good-Turing smoothed trigram language model was estimated. In language models that are based on written material only, a fixed context independent (i.e. unigram) probability was used for the interjections (estimated on broadcast news training data).

The ESAT decoder performs a single pass time synchronous beam search which results in real time recognition for the proposed task and (acoustic) modelling.

4.2. Spontaneous transcripts

Given the described CGN corpus, several subcorpora of spontaneous transcripts can be defined: as mentioned in the database description, on the one hand the spontaneous data in CGN consists of 3 parts (face to face dialogues, dialogues over telephone and broadcast data), and on the other hand CGN can be divided in Flemish and Dutch data. In order to define a suitable baseline spontaneous transcription corpus, different options were evaluated by

[1] Web site http://lands.let.ru.nl/cgn/ehome.htm

[2] Web site http://bach.arts.kuleuven.ac.be/fonilex

	Flemish		Dutch		both	
CGN face to face	1.0M	38.5%	2.1M	39.5%	*3.1M*	*37.4%*
CGN telephone	1.0M	40.0%				
CGN broadcast	0.6M	36.3%	1.3M	37.5%	1.9M	35.4%
CGN all 3	2.5M	36.1%			7.2M	35.1%
De Standaard only					33.3M	36.9%
De Standaard + CGN face to face					36.3M	34.7%
De Standaard + CGN all 3					40.5M	34.0%

Table 1: Spontaneous language models: training set size and WER (on the test set with Flemish broadcast data) given

recognition experiments on the test set. An overview of the results is given in table 1, both for language models based on spontaneous transcripts only, and for language models also including the *De Standaard* news text corpus.

From the table, we can conclude the following:

- using spontaneous data, language models based on 1.0M words or even less can be made that still give good results (comparable to the use of 33M words of news paper text).
- as expected, Flemish spontaneous broadcast data is best fitted to the task (namely the recognition of Flemish spontaneous broadcast audio). Telephone data seems to be worse than face to face data and was therefore not used in the final baseline.
- Dutch data doesn't fit as well as Flemish data (worse results for more data) but adding the Dutch data to the Flemish data still improves the result.
- adding news paper data to spontaneous data (or vice versa) improves the result.

It should be noted that using the CGN broadcast spontaneous transcriptions should be avoided: even though training and test data files are strictly separated, both contain news from the same time period and as the medium is the same (broadcast data) sentences are sometimes identical. Fortunately when including the news paper data, the difference between using only face to face data and using all 3 types of spontaneous speech is fairly small, for language models based on the spontaneous data only this is not the case. So as baseline spontaneous transcription corpus for the experiments in the next section, the CGN face to face component (including both the Flemish and the Dutch part) was selected.

5. Experiments and discussion

This section describes our experiments concerning the use of a large news paper text database for improving a spontaneous language model. The idea is to select the more *spontaneous* news paper texts from the large corpus and use them as language model training data. For the experiments in this section, the *Mediargus* database was used, consisting of over 1M of articles.

# words	random	unigram	trigram
350M	195.3		
100M	202.4	191.9	203.0
35M	212.9	195.3	224.3
10M	237.8	213.9	278.1
3.5M	263.8	253.2	372.4
350M	34.2%		
100M	35.5%	34.0%	34.7%
35M	35.9%	34.0%	34.9%
10M	37.3%	35.3%	36.9%
3.5M	37.7%	37.1%	39.2%

Table 2: Selecting written text for language modelling: perplexity (top) and WER given

In order to select articles that resemble spontaneous speech, a target language model is estimated on spontaneous speech. Texts then are ranked by their perplexity for this target language model, and texts with the lowest perplexity are selected for the language model training corpus. As target language model material the CGN broadcast data was used, as always excluding the recognition test set. Article selection based on both a unigram and a trigram language model was evaluated. With a bigram or a fourgram language model almost the same article ranking is found as with a trigram, so we didn't investigate this further. For reference, a random article selection was also investigated.

In tables 2 and 3 the results of the experiments can be found. For the first table, the language models are based on the selected written text material only, for table 3 the 3.1M words of CGN face to face dialogues were added to the language model training material. The first column always indicates the number of news paper words that are selected. Note that the result in table 3 when no text data is used (only the spontaneous data) is slightly different from the corresponding result in table 1 because different settings for the language model construction were used: the frequency cut-offs in this section were lower, so language models are larger for the same amount of training data.

Conclusions from both tables are almost the same:

- in case of the random selection the results improve when more text data is used, though there seems to

# words	random	unigram	trigram
350M	208.8		
100M	194.1	191.8	196.9
35M	185.7	182.5	195.7
10M	181.5	183.2	203.2
3.5M	186.5	198.2	218.2
0M	246.7		
350M	34.3%		
100M	34.9%	34.0%	34.4%
35M	34.4%	*33.4%*	33.8%
10M	35.2%	33.5%	34.8%
3.5M	35.0%	35.2%	35.4%
0M	37.0%		

Table 3: Selecting written text, language model training data includes spontaneous transcripts

be a saturation at 35M when the spontaneous data is added. The latter is due to the fact that using more data results in more detailed language model statistics but these statistics fit worse to the spontaneous test data than the statistics from the transcripts of spontaneous speech.

- for the selection based on N-grams, improvements by using more data stop at 35M, indicating that the statistics of additional (and according to the selection less spontaneous) data aren't of any help. When adding the spontaneous data, results even deteriorate by adding more than 35M words of written text.

- in WER, unigram selection is better than trigram selection, which in its turn is better than random selection (except for a too small training corpus size).

- compared to random selection, perplexities are better when using unigram selection (except again for a too small training corpus size) but clearly worse for trigram selection. As this differs from the situation concerning the WER, we can conclude that perplexity measures aren't a good indicator for the quality of a language model for recognition.

Overall we can conclude that the proposed technique of written text selection can improve the absolute WER by 0.9% from 34.3% (using all *Mediargus* data) to 33.4% (selecting 10% of the *Mediargus* data) while random selection deteriorates the results.

6. Conclusion

In this paper, a method was proposed to make use of a large database of written text in order to improve language models for spontaneous speech. The method selects appropriate texts from the large database in order to estimate the language model. It was shown that the method leads to a significant improvement over selecting random texts or selecting all data in the large database.

However the absolute WER for recognition of spontaneous speech in broadcast audio still is rather high (e.g. 33.4% for the investigated test set). Specific research towards disfluency handling in the language model (e.g. [5]) also results in significant but rather small improvements. It is supposed that larger improvements in large vocabulary spontaneous speech recognition will (or should) come from the acoustic level, or from an understanding component.

7. References

[1] V. Stouten, H. Van hamme, J. Duchateau, and P. Wambacq, "Evaluation of model-based feature enhancement on the AURORA-4 task," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 349–352.

[2] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, M. Pitz, and A. Sixtus, "The Philips/RWTH system for transcription of broadcast news," in *Proc. EUROSPEECH*, vol. II, Budapest, Hungary, Sept. 1999, pp. 647–650.

[3] J. Gauvain, L. Lamel, G. Adda, and M. Jardino, "Recent advances in transcribing television and radio broadcasts," in *Proc. EUROSPEECH*, vol. II, Budapest, Hungary, Sept. 1999, pp. 655–658.

[4] H. Yu, T. Tomokiyo, Z. Wang, and A. Waibel, "New developments in automatic meeting transcription," in *Proc. ICSLP*, vol. IV, Beijing, China, Sept. 2000, pp. 310–313.

[5] J. Duchateau, T. Laureys, and P. Wambacq, "Adding robustness to language models for spontaneous speech recognition," in *Proc. ISCA Workshop on Robustness Issues in Conversational Interaction*, Norwich, UK, Aug. 2004, paper 11 (4 pages). ISCA Archive, http://www.isca-speech.org/archive/robust2004.

[6] G. Adda, M. Jardino, and J. Gauvain, "Language modeling for broadcast news transcription," in *Proc. EUROSPEECH*, vol. IV, Budapest, Hungary, Sept. 1999, pp. 1759–1762.

[7] S. Robertson and K. Spark Jones, "Relevance weighting of search terms," *J. American Society for Information Science*, vol. 27, pp. 129–146, 1976.

[8] D. Biber, "A typology of English texts," *Linguistics*, vol. 27, pp. 3–43, 1989.

[9] J. Ponte and W. Croft, "A language modeling approach to information retrieval," in *Proc. ACM SIGIR*, 1998, pp. 275–281.

[10] J. Duchateau, K. Demuynck, and D. Van Compernolle, "Fast and accurate acoustic modelling with semi-continuous HMMs," *Speech Comm.*, vol. 24, no. 1, pp. 5–17, Apr. 1998.

[11] K. Demuynck, J. Duchateau, D. Van Compernolle, and P. Wambacq, "An efficient search space representation for large vocabulary continuous speech recognition," *Speech Comm.*, vol. 30, no. 1, pp. 37–53, Jan. 2000.