Use and Evaluation of Prosodic Annotations in Dutch

Jacques Duchateau, Tim Ceyssens, Hugo Van hamme

Katholieke Universiteit Leuven, Department ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
Jacques.Duchateau@esat.kuleuven.ac.be

Abstract

In the development of annotations for a spoken database, an important issue is whether the annotations can be generated automatically with sufficient precision, or whether expensive manual annotations are needed. In this paper, the case of prosodic annotations is discussed, which was investigated on the CGN database (Spoken Dutch Corpus). The main conclusions of this work are as follows. First, it was found that the available amount of manual prosodic annotations is sufficient for the development of our (baseline, decision tree based) prosodic models. In other words, more manual annotations do not improve the models. Second, the developed prosodic models for prominence are insufficiently accurate to produce automatic prominence annotations that are as good as the manual ones. But on the other hand the consistency between manual and automatic break annotations is as high as the inter-transcriber consistency for breaks. So given the current amount of manual break annotations, annotations for the remainder of the CGN database can be generated automatically with the same quality as the manual annotations.

1. Introduction

Prosody plays an important role in spoken communication between humans. However, for communication between human and machine, for instance in dialogue systems using spontaneous speech, the understanding and modelling of prosody mechanisms is still in its early stages. Even using nowadays large vocabulary dictation applications, in which the human tends to use more simple prosodic patterns, practical experience learns that the few errors in the recognition result often reflect inconsistencies with the prosody pattern that seem to be easily solvable, for instance inconsistencies with lexical stress.

So a lot of research is still to be done concerning the use of prosody to improve speech recognition. However in order to be able to investigate prosodic models, sufficiently large speech databases including prosodic annotations are needed. In the development of such database, an important question is if prosodic annotations can be generated automatically with sufficient accuracy, or if expensive manual annotations are needed. Even in the latter case, it may be advantageous (in terms of time spent by the transcriber) to use automatic annotations as a starting point and correct those manually.

In this paper, we investigate for a Dutch database if prosodic models can be developed that produce reliable prosodic annotations automatically, and how large an initial set of (manual) prosodic annotations should be to develop these prosodic models.

The structure of this paper is as follows. In the next section, the Dutch database is discussed, with a stress on the prosodic annotations in it. Then a section follows on the practical use of the database, and the tools needed to link the different information sources. Next the decision tree based prosodic models are described and the features used in it. The following section gives and discusses the experimental results, and finally some conclusions and directions for future work are given.

2. The CGN database

The CGN database (Corpus Gesproken Nederlands, Spoken Dutch Corpus [1]) was constructed between 1998 and 2003 in Flanders and in the Netherlands. It contains a total of 10 million words, or about 1000 hours of speech. The database consists of several components, with different types of audio: spontaneous speech and dictated speech, monologues and dialogues, standard recordings in a room, recordings over telephone, recordings of broadcast audio. In this study, we only use the Flemish part of the database, which is about one third of the total.

The CGN database contains several manually generated or checked annotations: orthographic transcriptions and part-of-speech tags for all acoustic data, and phonetic transcriptions, word level alignment, prosodic annotations and syntactic annotations for a smaller part of the data.

As manual prosodic annotations are quite expensive, they are available for only about 4% of the Flemish part of the data (125k words, corresponding to almost 15 hours of speech). In fact the annotations were made twice, by two different transcribers, and both annotations can be found in the CGN database.

For more information on the how and why concerning the prosodic transcriptions in the CGN database, the reader is referred to (Buhmann et al., 2002). Due to budgetary constraints (thus the use of non-expert transcribers) and given the envisaged database size, it was decided not to use a fine-grained labelling like ToDI [2], but a simpler, perceptually-based annotation as described in (Portele and Heuft, 1995).

The prosodic annotation consists of markers in the orthographic (graphemic) transcriptions both for syllables which carry the sentence accent (prominence) and for weak and strong breaks (prosodic boundary strength). The extra markers for unusual lengthening of sounds are ignored in this paper.

In the pilot study described in (Buhmann et al., 2002),

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

[2] Information on http://lands.let.kun.nl/todi

the inter-transcriber consistency (between the two prosodic transcriptions) was evaluated using Cohen's kappa coefficient [3]. For prominence, a kappa of on average 0.62 was found, and for break strength a kappa of 0.73, both indicating a substantial consistency.

3. Using the prosodic annotations in CGN

In order to develop prosodic models, a description of the audio data is needed with phoneme level segments, and for each segment it should be annotated if one of the prosodic events occurs or not. In the CGN database, this information is distributed over several annotations which require different tools to link them. In this section the used resources are described, and the methods to link them.

3.1. Basic annotations and resources

The following annotations and resources were used to develop the prosodic models:

- the orthographic transcription as available in the CGN database.
- the phonetic transcription, available for all data for which a prosodic transcription exists. If necessary, accurate phonetic transcriptions can also be generated automatically given the orthographic transcription as described in (Demuynck et al., 2002), using automatically generated multiple phonetic transcriptions for the words in the orthographic transcription, assimilation and other rules, and an alignment based on acoustic models to select the best fitting phonetic transcription given the audio.
- an automatic alignment of the phonetic transcription to the audio in order to know start and end point of all phonemes. These alignments will be generated for the CGN database using the sophisticated alignment system described in (Laureys et al., 2002). In this work however, a less accurate plain Viterbi alignment was used. The acoustic models needed for this alignment are estimated based on the dictated speech in CGN (a CGN component for which no prosodic annotations are available). Note that for telephone data, specific acoustic models are needed, therefore the prosodic annotations for telephone data were excluded for the experiments in this paper.
- the prosodic transcriptions: prominence and break strength are annotated on orthographic (graphemic) annotations (not in the phonetic transcriptions).
- a dictionary with (canonic) phonetic transcriptions which include the lexical stress of the word. For 6k frequent words, CGN copies this information from the Celex database [4]. For the 2k other words (compounds, proper names, etc.) CGN only offers an automatically generated canonic phonetic transcription without the lexical stress. Therefore this lexical stress was added manually for 2k words. However the lexical stress for words can probably be generated automatically in the same way and with the same accuracy as the canonic phonetic transcriptions themselves.
- for the annotation of syllable boundaries in the canonic phonetic transcriptions of the words, the situation in CGN is exactly the same as for the lexical stress. However for the experiment in this paper, this information was not used, instead (approximate) syllable boundaries were generated automatically using word-independent rules for separating consonant clusters.

3.2. Linking the annotations

In order to use the above information sources, the following links between the annotations are needed:

- a link between each word in the prosodic transcription and the corresponding word in the orthographic transcription. This is a 1-to-1 match (consistently followed throughout CGN), but the sentences are chunked differently and some parts of the transcriptions are (intentionally) lacking in the prosodic annotations. However, also given the available time markers, this link is easily made.
- a link between each word in the orthographic transcription and a string in the phonetic transcriptions. Again there is a consistently followed 1-to-1 match, and in this case the same chunks are used in both annotations. But one has to take into account the rules for cross-word degemination and linking phonemes. In some cases this results in words of which not a single phoneme is left, and this situation is difficult to detect and handle. Therefore the (few) chunks with these words were removed from the database.
- a link between each grapheme in the grapheme string for a word (which carries the prosodic markers) and the corresponding phoneme in phoneme string (where the prosodic markers are needed). To do this, first grapheme clusters that typically correspond to only one phoneme are turned into a single grapheme. Next the necessary links between grapheme and phoneme can be put in most cases using only the position of the grapheme and phoneme in the respective strings. So the identity of the grapheme and phoneme is not used except that it is checked that each prominence marker is put on a vowel phoneme. This may be surprising as a non-canonic, manually generated phoneme string for spontaneous speech is aligned to the standard orthography. Cases in which this alignment fails (mainly foreign names and abbreviations) are also removed from the database (this is about 1% of the database).
- a link between the canonical phonetic transcription which carries the lexical stress markers, and the realised phonetic transcription generated by hand on which the lexical stress markers are needed. Again an alignment was made based on the position of the phonemes in both strings and the fact the lexical stress markers should be put on vowel phonemes. This alignment works well for almost all words.

[3] See e.g. http://medcalc3000.com/Kappa_Cohen.htm

[4] Available from the LDC (Linguistic Data Consortium): http://www.ldc.upenn.edu

4. The prosodic model

Given the constructed database, decision tree based prosodic models for prominence and for break strength were developed using the C4.5 [5] software for decision tree construction and evaluation.

The basic features on which the prosodic models are based, are the segment duration (available in the automatic phoneme level alignments), the frame energy (the logarithm of the average squared sample value) and the pitch (fundamental frequency F0 evaluated with a home-grown pitch tracker). From the basic features, the used features are derived.

For the prominence model, the following derived features were calculated per syllable:

- NextSylShape. The shape of the next syllable's pitch contour. Classified in categories rise, fall, rise-fall, fall-rise by comparing the intitial, final, and mean values within the syllable.
- Ratio_ThisSylF0max_NextSylF0mean. The ratio of the maximum of the pitch contour in this syllable, to the mean of the next syllable's contour.
- Ratio_ThisSylF0max_PrevSylF0max. Similar ratio, but with the maximum of the previous syllable's contour as the denominator.
- Ratio_ThisSylF0max_ThisSylF0mean. Similar, but with this syllable's contour mean as the denominator.
- Ratio_ThisSylF0min_ThisSylF0mean. Similar, but with this syllable's contour minimum and this syllable's contour mean.
- ThisSyl_LexicalStress. A flag indicating whether this syllable carries a lexical stress.
- MND_ThisSylRhyme. Mean normalised duration (MND) of this syllable's rhyme. To calculate this and the two next features, statistics of phoneme lengths were recorded, i.e. means and variances for each phoneme. Then the actual length of the syllable/rhyme was compared to its expected length. The latter being estimated using the recorded means and variances.
- MND_PrevSyl. MND of previous syllable.
- MND_NextSyl. MND of next syllable.
- Vowel_energy. The mean energy of the vowel, normalised by dividing by the mean energy of the database file (recording session) of occurrence.
- SylSlope. The slope of the pitch contour over this syllable.

The importance of the pitch features arises from the fact that prominences and many other prosodic events are typically accompanied by certain pitch movements. The ratios try to reflect this. The lexical stress flag should be important because theoretically, prominence can only occur on lexically stressed syllables. A difference between the MND's and thus a locally changing speaking rate could also contain prosodic information, and the vowel energy is typically higher than normal in syllables which carry prominence.

For the break strength model, all features above are reused except for the lexical stress flag. The following extra features were used (all features are calculated after each word):

- Pause_duration. The length of the pause (silence) following the word. If there is no pause, this feature is obviously zero.
- W1Slope. The slope of the pitch contour over the last word. Calculated by applying linear regression.
- W2Slope. Similar, but over the last two words.

The importance of the pause feature is substantial: a strong prosodic break is typically accompanied by a silence. Some pitch ratios pertaining a wider range are now introduced, because pre-break prosodic phenomena are, on the average, expected to be spread more widely than prominence phenomena. The MND features are now at least as important as for prominences: a strong prosodic break is frequently accompanied by pre-boundary lengthening. This means that the last syllable, or even the last two syllables, are spoken more slowly than would be the case if they did not occur right before a break.

5. Experiments and discussion

Excluding telephone recordings (as mentioned above), 164 database files (typically recording sessions) in CGN contain prosodic annotations. For the test set, 16 files were selected randomly over the CGN components, resulting in 16k and 10k test cases for the experiments on prominence and on breaks respectively. The training set consists of the remaining files. For the experiments with varying training database size, the subsampling is done at the level of the training cases, not at the level of the files. In the experiments, the inter-transcription consistency is evaluated using Cohen's kappa coefficient (see section 2.).

5.1. Modelling prominence

Between the two manual prominence annotations, a kappa coefficient of 0.570 was found. Between the automatic prominence annotations and the manual ones, kappa coefficients of 0.337 and 0.430 were found, which are significantly worse.

Investigating a varying training database size, the results in figure 1 were found. The developed prominence models based on more than 10k training cases don't produce significantly different results.

So the developed prominence model is insufficiently accurate to produce annotations of the same quality and consistency as manual annotations, and using more manual annotations will not improve on that, at least for the proposed baseline modelling technique. However this doesn't mean that more sophisticated prominence models (for instance conditioning on the rhyme or even on the syllable) wouldn't produce better models, or wouldn't benefit from more manual annotations.

5.2. Modelling breaks

For the break case, a kappa coefficient of 0.703 was found between the two manual break strength annotations. Between the automatic break strength annotations and the manual ones, kappa coefficients of 0.721 and 0.730 were

[5] Available from http://www.cse.unsw.edu.au/∼quinlan/

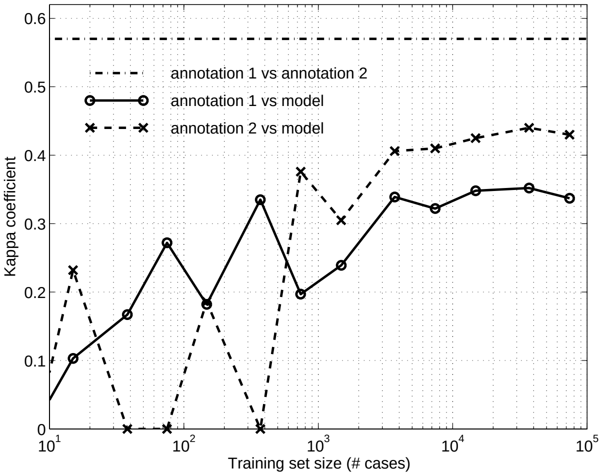

Figure 1: Consistency of prominence model

found. Given the test set size, all three results are statistically the same.

Figure 2 shows the results with a varying training database size. No significant improvements are found when a training database with a few thousand break cases is increased in size.

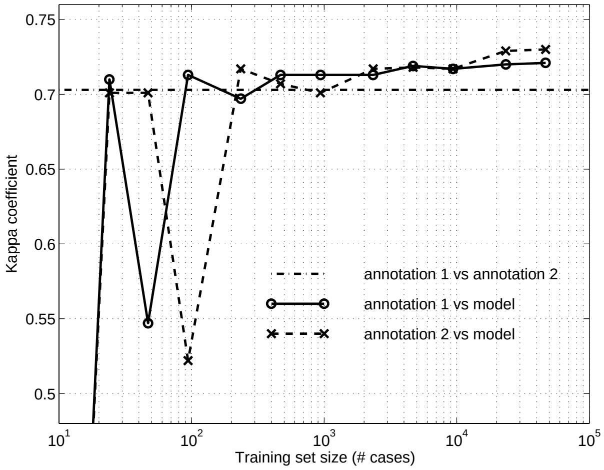

Figure 2: Consistency of break strength model

It can be concluded that for the perceptually-based break annotations in CGN a break strength model can be developed that is sufficiently accurate to produce annotations of the same quality and consistency as the manual annotations. Moreover a training database of few thousands of break cases was large enough to train this break strength model.

6. Conclusions and future work

In this paper the accuracy of prosodic models based on the CGN database was evaluated and compared to the accuracy of manual annotations.

A first conclusion is that the available amount of manual prosodic annotations in the CGN database is sufficient for the development of the proposed baseline prosodic models, a 5 times smaller training database would also suffice.

Secondly, comparing automatic and manual annotations, it was found that the developed prosodic models for prominence are insufficiently accurate to produce automatic prominence annotations that are as good as the manual ones. But on the other hand the consistency between manual and automatic break annotations is as high as the inter-transcriber consistency for breaks. So given the current amount of manual break annotations (or even a 5 times smaller amount), annotations for the remainder of the CGN database can be generated automatically with the same quality and consistency as the manual annotations. Note that in order to generate those automatic prosodic transcriptions, only an orthographic transcription (available in CGN) is needed. As explained in section 3.1., phonetic transcriptions, phoneme level alignment and lexical stress markers can be generated automatically.

In order to improve the prominence models, it should be investigated whether specific models for the different database components could help. An other option is to try more sophisticated models, e.g. through conditioning on rhyme or syllable identity. Note that both options may raise the need for a larger manually annotated database. Moreover even with the current models, it may be advantageous to use automatic prominence annotations in order speed up the development of manual ones. But this should be checked in practice.

Acknowledgements

This research work was supported by the Flemish Government through IWT-Vlaanderen. Tim Ceyssens is a Research Assistant of the Fund for Scientific Research - Flanders (Belgium) (F.W.O.-Vlaanderen).

7. References

Buhmann, Jeska, Johanneke Caspers, Vincent van Heuven, Heleen Hoekstra, Jean-Pierre Martens, and Marc Swerts, 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken Dutch corpus. In *Proc. 3rd International Conference on Language Resources and Evaluation*, volume III, pp. 779–785. Las Palmas de Gran Canaria, Spain.

Demuynck, Kris, Tom Laureys, and Steven Gillis, 2002. Automatic generation of phonetic transcriptions for large speech corpora. In *Proc. 7th International Conference on Spoken Language Processing*, volume I, pp. 333–336. Denver, U.S.A.

Laureys, Tom, Kris Demuynck, Jacques Duchateau, and Patrick Wambacq, 2002. An improved algorithm for the automatic segmentation of speech corpora. In *Proc. 3rd International Conference on Language Resources and Evaluation*, volume V, pp. 1564–1567. Las Palmas de Gran Canaria, Spain.

Portele, Thomas and Barbara Heuft, 1995. Two kinds of stress perception. In *Proc. 13th International International Congress of Phonetic Sciences*, volume I, pp. 126–129. Stockholm, Sweden.