Automatic Phonemic Labeling and Segmentation of Spoken Dutch

Kris Demuynck, Tom Laureys, Patrick Wambacq, Dirk Van Compernolle

Katholieke Universiteit Leuven – ESAT
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{kris.demuynck,tom.laureys,patrick.wambacq,dirk.vancompernolle}@esat.kuleuven.ac.be

Abstract

The CGN corpus (Oostdijk, 2000) (Corpus Gesproken Nederlands/Corpus Spoken Dutch) is a large speech corpus of contemporary Dutch as spoken in Belgium (3.3 million words) and in the Netherlands (5.6 million words). Due to its size, manual phonemic annotation was limited to 10% of the data and automatic systems were used to complement this data. This paper describes the automatic generation of the phonemic annotations and the corresponding segmentations. First, we detail the processes used to generate possible pronunciations for each sentence and to select to most likely one. Next, we identify the remaining difficulties when handling the CGN data and explain how we solved them. We conclude with an evaluation of the quality of the resulting transcriptions and segmentations.

1. Introduction

The CGN corpus (Oostdijk, 2000) (Corpus Gesproken Nederlands/Corpus Spoken Dutch) is a large speech corpus of contemporary Dutch as spoken in Belgium (3.3 million words) and in the Netherlands (5.6 million words). The audio data are enriched with several transcriptions and segmentations: orthography, part-of-speech tags and automatically generated phonemic transcriptions with corresponding segmentations are available for the whole corpus. Syntactic annotations, manually verified phonemic transcriptions and word segmentations are provided for 10% of the data (the core corpus), while prosodic labeling was carried out on 5% of the data. In this paper, we describe the methods and techniques used to create the automatically generated phonemic transcriptions and segmentations for the Belgian (Flemish) part of the database.

First, we outline the process of automatically generating phonemic transcriptions with their corresponding segmentations. Next, we identify the remaining difficulties when handling the CGN data and explain how we solved them. We conclude with an evaluation of the quality of the resulting transcriptions and segmentations.

2. From Orthography to Pronounciation

The orthographic annotations in CGN are enriched with various markers providing the following information: [⋆v] foreign word, [⋆a] incomplete (broken-off) word, [⋆x] ill-understood word (educated guess from the transcriber), [⋆u] onomatopoeia or mispronunciation, [⋆z] Dutch word pronounced with a strong regional accent, and [⋆d] dialect word. The presence of capitals and digits provides further information on the function of the word and its possible pronunciations. Abbreviations are written in all capitals with no dots in between, while numbers that are part of abbreviations (e.g. BBC1) are transcribed with digits. Capitals are also used to mark proper nouns (e.g. ‘New York’) or titles of books, movies, songs and so on (e.g. ‘The⋆v Deer⋆v Hunter⋆v’). Note that foreign proper nouns are not marked with a ⋆v, whereas titles are. Sentences end with a punctuation mark, but do not start with a capital.

Based on the orthographic input, all plausible pronunciations of the sentences are automatically generated and the acoustically best matching phonemic sequence is selected. To obtain plausible pronunciations for the words, the following techniques and resources were used:

Lexicon lookup: Fonilex (Mertens and Vercammen, 1998) provides multiple phonemic transcriptions for all frequent standard Dutch words. For the foreign words we draw on Comlex (English), Celex (German) and Brulex (French). If a foreign word is part of more than one of these lexica, the different phonemic realizations are put in parallel since the orthography does not specify which foreign language is used. The same holds for capitalized words (e.g. ‘Hamburg’ which may either be pronounced in a Dutch, German or English fashion). Furthermore, specific lexica were made for the most frequent proper nouns (5892 entries), interjections, frequently used dialect words and items not covered in one of the other lexica (982 entries).

Compounding, derivation and inflection: As Dutch is a morphologically productive language, lexica in itself cannot cover all possible word forms. The pronunciation of non-trivial compounds and derivations is found by decomposing the word into its basic constituents, concatenating their pronunciations and applying a set of assimilation rules. In our approach, all decompositions possible based on pure orthographic constraints are pursued, i.e. no morphotactic constraints are imposed. So some degree of overgeneration is introduced (e.g. ‘rijstroken’ → ‘rij’ + ‘stroken’ / ‘rijst’ + ’roken’). This overgeneration rarely resulted in new pronunciation variants and even showed to be useful for handling Dutch proper nouns and mispronunciations.

Abbreviations and digits: Abbreviations are phonemically transcribed as the concatenation of the constituent letter word transcriptions. In case the abbreviation –converted to lower case– maps to an existing word, the corresponding word pronunciation is added as well. Frequently occurring exceptions (e.g. NATO) are added to one of the specific lexica. The pronunciation of numbers inside the abbreviations is solved with a rule-based system.

Broken-off words: Broken-off words (with broken-off orthography) are searched in a grapheme-phoneme aligned version of the Fonilex database and the pronunciations for all matching entries are put in parallel.

Strong regional accents: Starting from the standard Dutch pronunciations, a set of context-dependent rewrite rules are applied in order to generate a large number of plausible dialect pronunciation variants (cf. infra: assimilation).

Grapheme-to-phoneme system: A grapheme-to-phoneme (g2p) system was developed as a fall-back. The g2p system is based on the Induction Decision Tree (ID3) mechanism (Pagel et al., 1998) and trained on the Fonilex database. More information on the configuration of the g2p system is given in (Demuynck et al., 2002).

Assimilation: In continuous speech, phonemes at word boundaries influence each other. These cross-word phenomena (assimilation, degemination, inserted linking phonemes, etc.) are handled by a set of rewrite rules of the form: phoneme sequence c (possibly empty) in the context $l \cdot c \cdot r$ is or can also be pronounced as c'. These rules are internally applied to the complete sentence by our speech recognition system (Demuynck, 2001), resulting in a compact pronunciation network. The set of rules used are a subset of the rules in the Fonilex database (most word-internal assimilation rules also operate across word boundaries), extended with rules found in other resources (Verschaeren and Van Compernolle, 1995).

For the Flemish data in CGN, 68% of the words in the word list (92540 entries) are directly covered by one of the lexica. New compounds and inflections derived from the Fonilex lexicon, broken-off words, and abbreviations account for 22.5%, 3.5% and 1% repectively. The remaining words are handled by the g2p system and are distributed as follows: proper nouns (1.6%), onomatopoeia and mispronunciations (0.9%), dialect words (0.5%), foreign words (0.4%), and new words or unhandled derivations (1.6%).

3. Automatic Alignment: Viterbi versus Forward-Backward

Once a pronunciation network is generated for every sentence, the transcription matching best with the speech signal is selected automatically. All phonemic alternatives are acoustically scored in a single pass (Viterbi alignment) through our speech recognition system using context-dependent (within- and cross-word) phoneme models and the most probable one is retained. More details on the recognition system and how it handles pronunciation networks can be found in (Demuynck, 2001). Details on the acoustic models will be given in section 4.

For the segmentation, one can either use the alignment provided by the Viterbi pass (maximum likelihood assignment of speech to phonemes) or one can opt for an additional Forward-Backward pass which optimizes the boundaries between phonemes in a least squares sense.

3.1. Viterbi Segmentation

The Viterbi algorithm returns the single best path through the model given the observed speech signal x_1^T:

$$s_1^T = \arg\max_{s_1^T \subset S} \prod_{i=1}^{T} f(x_i \mid s_i) p(s_i \mid s_{i-1}) \ ,$$

with s_1^T a sequence of HMM states (one state for each time frame) which is consistent with the sentence model S, and T being the number of time frames. Thus, the Viterbi algorithm results in the segmentation which reaches maximum likelihood for the given feature vectors.

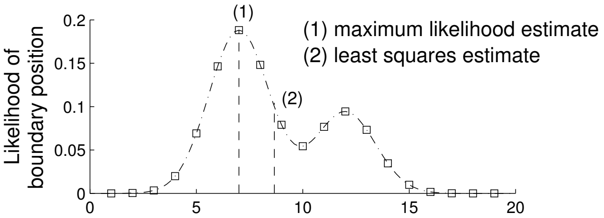

Figure 1: Viterbi and Forward-Backward boundaries

3.2. Forward-Backward Segmentation

The Viterbi algorithm only provides us with an approximation of the optimal boundary position. This is illustrated in figure 1. The Viterbi algorithm generates the boundary corresponding to (1), whereas the optimal boundary in a least squares sense matches with (2). Assuming that humans more or less follow the same likelihood distribution when placing boundaries, solution (2) will be closest to the manual segmentation on average.

To find the best possible estimate of the boundary in a least squares sense the probability function of each boundary must be calculated:

$$P(b|S, x_1^T) = \frac{f(x_1^b|S_l) f(x_{b+1}^T|S_r)}{f(x_1^T|S)} \ ,$$

with

$$f(x_a^b|S_x) = \sum_{s_a^b \subset S_x} \prod_{i=a}^{b} f(x_i|s_i)^{1/\beta} p(s_i|s_{i-1})^{1/\beta}$$

In the above equations, sentence S is divided in part S_l left and part S_r right of the boundary of interest. The extra parameter β compensates for the ill-matched assumption made by HMMs that the observations x_i are independent. The optimal value for β in our experiments was 10, but its exact value was not at all critical. The same compensation factor can be found in recognition systems (Demuynck, 2001) as well as in confidence scoring of recognized words (Wessel et al., 2001) for balancing the contribution of acoustic and language model. The Forward-Backward algorithm allows for an efficient calculation of the density functions for all boundaries in a sentence. Given the probability density function of each boundary, the least squares estimate $E\{b\}$ equals:

$$E\{b\} = \sum_{b=1}^{T} P(b|S, x_1^T) \, b$$

4. Automatic Channel and Mode Selection

To allow for efficient batch processing of all CGN-data, a number of difficulties have to be coped with. First, CGN contains audio data from a wide variety of sources (see table 1), resulting in recordings at office (16kHz) and telephone (8kHz) quality of both polished and spontaneous speech. Since all polished speech is recorded at 16kHz, and since the additional bandwidth provided by 16kHz recordings of spontaneous speech has little effect on the accuracy of our automatic system, we opted to distinguish between two modes only: 16kHz polished speech on the one hand and all remaining speech down-sampled to 8kHz on the other hand. For both modes, a specific Hidden Markov

a	spontaneous conversations (face-to-face)
b	interviews with teachers of Dutch
c	spont. telephone dialogs recorded via a switchboard
d	spont. telephone dialogs recorded locally (tel. tap)
e	empty for the Flemish part of the corpus
f	interviews/discussions/debates (broadcast)
g	(political) discussions/debates/meetings (non-broadcast)
h	lessons recorded in a classroom
i	live (e.g. sport) commentaries (broadcast)
j	news-reports/reportages (broadcast)
k	news (broadcast)
l	commentaries/columns/reviews (broadcast)
m	ceremonious speeches/sermons
n	lectures/seminars
o	read speech (books on tape)

Table 1: The different components of CGN

	a	b	c	d	f	g	h	i	j	k	l	m	n	o
ins	5.2	2.3	5.9	8.5	2.5	1.1	2.2	1.6	1.3	1.2	1.3	1.4	2.4	1.1
del	1.9	1.6	1.6	1.6	1.7	1.5	1.6	1.5	1.3	1.5	1.4	1.7	1.7	1.2
sub	6.3	4.7	6.9	7.8	4.5	3.4	3.6	4.0	3.3	2.8	2.6	4.2	3.6	2.4

Table 2: Deviations (in percent) between automatic and manual phonemic transcription for all components in CGN

type	freq.	details & relative importance	
ins.	1.1%	n: 53.6% ə: 19.8% j/w: 6.9% l/r: 3.7% t/d: 3.4%	
del.	1.2%	n: 28.2% h: 23.9% ə: 20.3% t/d: 10.7% j: 4.5%	
sub.	2.4%	inter vowel:	50.5%
		long → corresponding short vowel:	18.8%
		short → corresponding long vowel:	7.6%
		ə ↔ e, E, I :	14.9%
		inter consonant:	44.1%
		unvoiced → corresponding voiced:	15.0%
		voiced → corresponding unvoiced:	11.3%
		nasals (n, m, ŋ):	9.4%

Table 3: Analysis (comp. o) of the most frequent errors

Model was created. During the final handling of the data, the speech mode for a given audio fragment is determined automatically: both acoustic models run in parallel and the model which reports the best acoustic match is selected. To allow a direct comparison of the acoustic scores, the state likelihoods for each frame were normalized as follows:

$$f'(x \mid s_i) = f(x \mid s_i) / \sum_k P(s_k) f(x \mid s_k)$$

with $P(s_k)$ the a priori (uni-gram) probability of each state. See (Demuynck, 2001), page 43 for more details on the rationale behind this normalization.

A second difficulty is the presence of both mono and stereo audio files. For stereo recordings, the dominant channel (left or right) for a given audio fragment and speaker is selected before the speech mode is determined. For this channel detection, the 8kHz acoustic model is run on speech from both the right and the left channel, and the channel providing the best match is selected.

Finally, the acoustic models have to be adapted to the specific task. When training acoustic models, the assimilation processes are usually modeled implicitly by making context dependent models, i.e. not the phonemes but the phonemes given their left and right contexts (phonemes) are modeled. For our purpose –the labeling of CGN data– such implicit modeling of the assimilation processes would result in inaccurate phonemic labeling, and could possibly conflict with our explicit set of assimilation rules. Hence, the same context-dependent rewrite rules as used with the automatic labeling of the CGN data are activated during training, assuring maximal consistency between the acoustic models and their final use for labeling. Assigning speech to a certain mode (model) was done manually: data of read-aloud books and newspaper texts, augmented with a small set of manually selected recordings from lessons, lectures and debates in order to have examples of speaker noises and non-understood words, were used to train the 16kHz polished speech model. The same data was also down-sampled to 8kHz and used to train the second acoustic model. Detection of the dominant channel for each speaker was done in advance with existing acoustic models.

The telephone recordings and downsampled data from some of the other components may be better suited to train the 8kHz spontaneous speech model. However, the large number of preprocessing steps needed to make this data suitable for training acoustic models (e.g. removing all data with overlapping speech from two or more speakers, or with excessive amounts of background noise) prevented us so far from creating these models. A later update of the CGN-corpus may provide new automatic phonemic transcriptions and segmentations if such 'matched' 8kHz model provides substantially better results.

5. Evaluation

5.1. Channel and Mode Selection

We evaluated the 8/16kHz detection by manually checking whether each speaker in component k was recorded in the studio or over a telephone line. This selection proved to be error free. Also the channel selection in stereo files seemed to work flawlessly, except for some telephone recordings with substantial crosstalk between the two channels. The selection of polished or spontaneous speech could not be evaluated, since both acoustic models were trained on polished speech only.

5.2. Automatic Phonemic Transcriptions

The automatic phonemic transcription was evaluated by counting the number of insertions, deletions and substitutions with respect to a hand-checked reference transcription. The reference transcription was produced by a trained phonetician who corrected a baseline transcription generated by a g2p system different from the one described in section 2. Table 2 summarizes the results, while table 3 gives an analysis of the most frequent errors for component o. The other components show very similar patterns.

A detailed study of the contexts in which these errors occur showed that not every deviation was a mistake on the side of the automatic system. For example, the automatic transcription typically incorporates more connected speech effects than its manual counterpart. This might be due to the fact that human transcribers, having to work at

comp.	Viterbi			Forward-Backward		
	35ms	70ms	100ms	35ms	70ms	100ms
a	26.8%	15.2%	11.4%	26.3%	14.3%	10.7%
b	16.5%	8.0%	5.6%	15.6%	7.4%	5.2%
c	22.4%	8.9%	5.2%	21.9%	8.1%	4.6%
d	23.1%	10.3%	6.4%	22.3%	9.3%	5.6%
f	12.8%	5.7%	4.0%	11.5%	5.1%	3.6%
g	12.2%	4.7%	2.8%	10.6%	3.4%	1.8%
h	15.1%	5.8%	3.3%	13.9%	5.1%	2.8%
i	11.7%	4.9%	3.1%	9.9%	3.9%	2.4%
j	8.0%	2.5%	1.4%	6.3%	1.9%	1.1%
k	8.7%	2.1%	1.1%	6.7%	1.5%	0.9%
l	8.3%	2.4%	1.3%	6.4%	1.9%	0.9%
m	15.2%	6.2%	3.7%	11.7%	4.6%	3.1%
n	14.8%	5.8%	3.3%	14.2%	4.9%	2.6%
o	8.8%	1.7%	0.5%	7.1%	1.1%	0.4%

Table 4: Freq. counts of deviations between automatic and manual word segmentations for all components in CGN

a considerable speed, sometimes overlook these phenomena not present in the base transcription they were offered. Some examples are: [1] ə-insertion in non-homorganic consonant clusters in code position (‘kalm’ /kAləm/) (Booij, 1995), [2] homorganic glide insertion between vowels (‘die een’ /dijən/), or [3] n-deletion due to nasal assimilation (‘onmacht’ /ɔmɑxt/). Similarly, phenomena already applied in the base transcription such as syllable-final n-deletion were not always undone when not present.

The errors that unambiguously correspond to an error in the automatic system are mainly due to common dialect phenomena on the word level not described in the dictionary (e.g. deletion of the final /t/ in frequent words like ‘dat’ and ‘wat’), and low signal quality or overlapping speech. In fact, when looking at those components with mainly high quality polished speech (j, k, l, and o), the amount of disagreement between automatic and manual transcription is very close to what we obtained when comparing two manual annotations made by different phoneticians.

5.3. Word Alignment

The automatic segmentations were evaluated by counting the number of word boundaries for which the deviation between automatic and manual segmentation exceeded thresholds of 35, 70 and 100ms. Manual segmentation started from an automatic segmentation produced by the Viterbi algorithm using acoustic models trained on an older Dutch database. The persons that made the manual segmentations were instructed to position boundaries so that each word would sound acoustically acceptable in isolation (Martens et al., 2002). Shared phonemes at the boundary (e.g. he is_sad) were split in the middle, except for shared plosives (e.g. stop_please), which were isolated altogether. Noticeable pauses ($>$ 50 ms) were segmented in the same way as words, thus producing empty chunks.

We evaluated both the Viterbi and Forward-Backward segmentation based on automatically generated phonemic transcription. As can be seen from the results in table 4, the forward-backward method outperforms the Viterbi approach on all components, giving a 15% reduction in error counts on average. A detailed analysis showed that more than half of the remaining 35 msec deviations in the automatic segmentations are transitions to and from noise and transitions to unvoiced plosives (e.g. 34.5%, 12% and 15% respectively for component o). Since these boundaries also show large variation between the corresponding manual segmentations of different correctors, we cannot expect an automatic system to give more consistent results.

6. Conclusions

We described the system used for the automatic generation of phonemic transcriptions and segmentation for the Flemish part of the CGN-corpus. First, different techniques are applied for the generation of all plausible pronunciation variants for a given orthographic transcription. Next, a speech recognition system is employed for selecting the best matching transcription with corresponding segmentation. Analysis of the differences between automatically generated annotations and those made by humans showed performance levels close to that of human transcribers for the ‘high quality polished speech’ part of the database. We also observed that the automatic system was not always to blame: humans make mistakes too, e.g. due to tiredness and loss of concentration, phenomena which never trouble automatic systems. For the ‘spontaneous speech’ parts in CGN, the results obtained by the automatic system are still up to standard, but the gap between the automatic system and human transcribers widens.

7. Acknowledgments

The research described in this paper was funded by AWI, project CGN (http://lands.let.kun.nl/cgn/), and by IWT in the GBOU program, project FLaVoR: Flexible Large Vocabulary Recognition (Project number 020192). (http://www.esat.kuleuven.ac.be/∼spch/projects/FLaVoR/).

8. References

Booij, G.E., 1995. *The Phonology of Dutch*. Oxford: Clarendon Press.

Demuynck, K., 2001. *Extracting, Modelling and Combining Information in Speech Recognition*. Ph.D. thesis, K.U.Leuven, ESAT.

Demuynck, K., T. Laureys, and S. Gillis, 2002. Automatic generation of phonetic transcriptions for large speech corpora. In *Proc. ICSLP*, volume I. Denver, U.S.A.

Martens, J.-P., K. Demuynck, D. Binnenpoorte, R. Van Parys, T. Laureys, W. Goedertier, and J. Duchateau, 2002. Word segmentation in the Spoken Dutch Corpus. In *Proc. LREC-2002*, volume V. Las Palmas, Canary Islands.

Mertens, P. and F. Vercammen, 1998. FONILEX manual. Technical report, K.U.Leuven – CCL. http://bach.arts.kuleuven.ac.be/fonilex/.

Oostdijk, N., 2000. The Spoken Dutch Corpus. *The ELRA Newsletter*, 5(2):4–8. http://lands.let.kun.nl/cgn/.

Pagel, V., K. Lenzo, and A.W. Black, 1998. Letter to sound rules for accented lexicon compression. In *Proc. ICSLP*, volume I. Sydney, Australia.

Verschaeren, K. and D. Van Compernolle, 1995. The phonological rules of Dutch. Internal Report PSI-SPCH-95-10, K.U.Leuven, ESAT.

Wessel, F., R. Schlüter, K. Macherey, and H. Ney, 2001. Confidence measures for large vocabulary speech recognition. *IEEE Trans. on SAP*, 9(3):288–298.