Towards Automatic Word Segmentation of Dialect Speech

Eric Sanders[1], Andrea Diersen[1], Willy Jongenburger[2], Helmer Strik[1]

[1]Radboud University Nijmegen, the Netherlands
[2]Meertens Instituut, the Netherlands
[E.Sanders,A.Diersen,H.Strik]@let.kun.nl, Willy.Jongenburger@meertens.knaw.nl

Abstract

This paper is about the creation of a digital dialect database, and the focus is on automatic word segmentation. Automatic word segmentation has been studied by several research groups during the last two decades. However, the task we are faced with differs in several respects from previous ones. For instance, in our case we are dealing with recordings of interviews containing spontaneous dialect speech and 'enriched' (quasi-phonetic) orthographic transcriptions (instead of 'normal' orthographic transcriptions, which are usually available). Furthermore, the nature of the task requires that the word segmentation procedure can be adapted for each interview.

1. Introduction

The Meertens Institute in Amsterdam [16] has a long history in dialect research, and for this research several dialect databases have been collected. One of these databases contains about 660 interviews that were recorded in various regions of the Netherlands during the second half of the last century. The dialects present in this database differ substantially from each other and from standard Dutch [11]: they have different phonological systems, and there are also lexical and syntactical differences (see, e.g., the examples in Tables 1 and 3). Both the speech signals and the orthographic transcriptions were originally available only in analog form: speech on magnetic tapes, and transcriptions (typed with a typewriter) on paper. In order to better preserve the material, speech signals and part of the transcriptions have been digitized. The next step, in order to increase the accessibility of the data, is to make a link between speech signals and orthographic transcriptions in terms of a word segmentation and alignment. This is the goal of the multimedia dialect database project (MuMDiD) that started recently [18]. In this pilot project, a word segmentation and alignment have to be produced for part of the database.

Automatic word segmentation has been studied by different research groups over the last two decades, see e.g. [13,14,10,6,15]. Angelini et al. [1] concluded that the task is more difficult if one has to start from an orthographic transcription instead of from a phonetic transcription. When only an orthographic transcription is available, pronunciation modeling can improve the performance of the automatic segmentation [3]. There have been very few studies on automatic speech recognition of dialect speech [7,5,2], and, as far as we know, none about automatic word segmentation and alignment of dialect speech.

Although automatic word segmentation has been studied before, the task we are faced with differs in several respects from previous ones. Some of these differences are briefly mentioned in this section, more details are provided below.

First of all, this database contains dialect speech and not standard language. Furthermore, this database has been collected over a long period, more than 30 years, and different sets of equipment (tape recorders, microphones, etc.) have been used over the years. The signals were usually recorded on magnetic tapes at the home of one of the informants. These tapes have been stored for many years (sometimes almost 50 years), and have been digitized only recently. The transcriptions were made by many different transcribers over a long period of time. Although the transcribers received some instructions, it is certainly not the case that there was a well-described transcription protocol that was used by all transcribers. Given all these differences between dialects, recordings, transcribers, and transcriptions, it will probably be necessary to optimise the segmentation procedure for each interview. All these differences between this task and previously studied tasks raise some new research issues. It is the goal of this paper to make an inventory of these issues, explain how they arose, and how they can be approached.

In the next section, the material of the dialect database as well as the training material for the automatic word segmentator is described. The third section will explain the methods to be used and the experiments that will be conducted to produce the word segmentation.

2. Material

2.1. Dialect speech

The current dialect database was collected between the early fifties and the early eighties of the twentieth century. The material consists of interviews and multi-party conversations. A contact person went to his or her native region to have an informal talk with one or more dialect speakers, usually at the home of one of the informants. These talks were more like free conversations than a question-answer dialog. There were no restrictions on the topic of discussion. In most cases the speakers were invited to talk about old customs and traditions. Farming and related issues were also often discussed. The recordings contain between 2 and 6 speakers, and informal listening to a small number of interviews revealed that there is a substantial amount of cross-talk.

In total, there are 660 recordings with a duration of about 480 hours. In the pilot project MuMDiD, 50 hours of speech (about 80 recordings) will be segmented. The interviews in this pilot project are selected in such a way that they are balanced according to dialect, gender, and number of speakers. In Figure 1, the different dialect groups in the Netherlands are shown [8]. Most of these regions will be covered in this pilot project. In Table 1 some examples of dialect speech are given.

Figure 1. *Dialect distribution in the Netherlands*

Language/ dialect	Example/translation
Aalten	Moar dan deden ze dat zo an banen ôver de dèle leggen
Standard Dutch	Maar dan legden ze dat zo in banen over de deel
English	But then they laid it in the stable in rows
Leiden	Alles *mos* so gekoowp moowgelijk.
Standard Dutch	Alles moest zo goedkoop mogelijk
English	Everything had to be as cheap as possible
Hoog Blokland	Toen eh *moes* ik bij ginneraol Snijders komme
Standard Dutch	Toen, eh, moest ik bij generaal Snijders komen
English	Then, err, I had to go to general Snijders
Steenbergen	Ik *most* driehonderd gulde trekke
Standard Dutch	Ik moest driehonderd gulden pakken
English	I had to take three hundred guilders

Table 1. Example sentences from different dialects with standard Dutch and English translation. The last three examples, contain different variations of the word 'moest' (had to).

For all interviews, 'enriched' (quasi-phonetic) orthographic transcriptions are available. For words that are pronounced the same way as in standard Dutch, the conventional Dutch spelling is generally used. For words that in Dutch are pronounced differently or do not exist at all, a transcription that reflects the pronunciation as closely as possible is produced. For non-standard Dutch sounds, special symbols are introduced. For most of the interviews, a short transcription convention is present containing a brief description of these special symbols. An example is provided in Table 2, containing the transcription convention of the dialect of Aalten (see the first example in Table 1). However, since no well-described transcription protocol exists, the way in which 'normal' symbols are used to transcribe pronunciations that differ from standard Dutch (see, e.g., in Table 1: Moar, gekoowp, moowgelijk, and ginneroal) differs between transcribers. Another inconsistency concerns, e.g., the use of the 'n' at the end of words. In Table 1, it can be observed that sometimes the 'n' at the end of a word is not transcribed, probably reflecting that the 'n' was not pronounced. However, this transcription convention was not consistently used by all transcribers.

Symbol	Pronunciation
è	like 'e' as in French 'la mere'
ô	like 'o' as in 'pot'
oa	like 'o' as in English 'more'
ij	like 'i' in 'pit' followed by 'j'

Table 2. Special pronunciation symbols from a recording of the dialect spoken in Aalten.

The original transcriptions were typed on paper. They were digitized by scanning and optical character recognition (OCR). After OCR all transcripts were manually checked and corrected by comparing the transcripts after OCR with those after scanning. The resulting transcripts contain some errors and 'blanks'. Some errors were already present in the original transcripts (typed on paper), and some errors were introduced by the digitization process (i.e., those that were not manually corrected). Because there is no such thing as a finite lexicon of correct forms, there is no way to detect typing and OCR errors automatically. The original analog transcripts contain some 'blanks', denoted as "…", for parts of the utterances that could not be transcribed. The digital transcripts contain some extra 'blanks', denoted by #'s, for parts of the original transcripts that could not be deciphered by the manual corrector, mainly because words were re-typed over incompletely erased typos.

2.2. Training material for phone models

The segmentation will be carried out with a speech recogniser based on phone models. The phone models will be trained on material taken from the Spoken Dutch Corpus (CGN: "Corpus Gesproken Nederlands") [19,4]. The CGN contains recordings of different speech styles: e.g., radio broadcasts, speeches, meetings, spontaneous conversations, telephone dialogues, and read speech. All speech is standard Dutch. Some speech was pronounced with an accent, but dialect speech is not included. All utterances have been orthographically transcribed by hand. A broad phonemic transcription was automatically derived for the whole corpus. For one million words (about 100 hours of speech) the phonemic transcriptions were manually checked, the transcriptions and the speech signal were time aligned automatically, and this alignment was checked manually (at the word level).

We will experiment using different sets of material to train the phone models. From the CGN, we will take speech material that resembles the type of speech used in the dialect database as much as possible, such as interviews and spontaneous conversations. In addition, we will experiment with using small amounts of dialect speech (taken from the current dialect database), e.g., for each dialect region, to adapt the phone models.

3. Method

The whole word segmentation procedure will consist of the following stages:

1. Chunking
2. Grapheme-to-phoneme conversion
3. Word segmentation
4. Quality control

3.1. Chunking

For each interview there is one speech file. The length of the speech files varies from 20 to 40 minutes. Obviously, segmentation should not be carried out directly on such long files, and thus chunking is needed. Chunking is done by students trained for this job, using the program *Praat* [20].

Praat allows one to listen to the speech signal and to view it on the screen together with the corresponding transcription. The students are instructed to place chunk boundaries in the speech signal about every 5 seconds, and to assign the correct part of the transcription to it (see Figure 2). The transcription of each speaker is in a separate tier. Boundaries are placed per speaker tier and preferably in natural speech pauses.

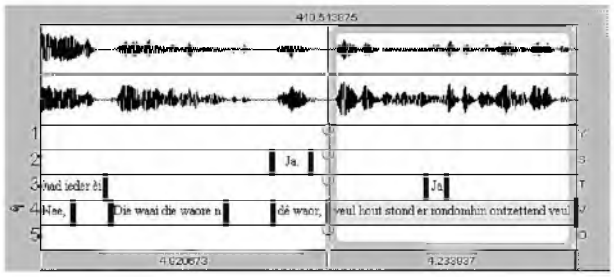

Figure 2. *Praat window, with which the chunking is done*

3.2. Grapheme-to-phoneme conversion

Deriving a phone transcription is quite a challenging task, since these transcriptions must be derived from the 'enriched' orthographic transcriptions (see section 2.1) of different dialects made by several transcribers over a long period, and the transcriptions contain inconsistencies, errors, and 'blanks' (see also section 2.1). Furthermore, there is a large amount of pronunciation variation between dialects. In order to illustrate this, examples of the same sentence in four different dialects are presented in Table 3 (taken from [21]).

Region	Example
<English>	And he stood up and returned to his father.
<standard Dutch>	En hij stond op en ging naar zijn vader terug.
Groningen	En hai ston op en ging noar zien voader tou.
Gelderland	En hie stung op en gung weer nur z'n vao.
Limburg	En haer sjting òp en ging nao z'ne vajer truuk.
Friesland	En hy stúech op en keersde wierom nooi syn heit.

Table 3. Examples of the same sentence in English, standard Dutch, and four different dialects.

Important research issues in the current project are how to convert the enriched orthographic transcription to a phone transcription, which set of phone symbols should be used (e.g., is it possible to use the same symbol set for all dialects), and what kind of acoustic models should be employed. The latter aspect is discussed in section 3.3.

Given the differences in dialects, transcribers, and transcription conventions, there are substantial differences between transcriptions. Therefore, in order to obtain the best results, it is necessary to use a grapheme-to-phoneme converter that can be optimised for each interview (or group of interviews with similar characteristics). This is one of the main reasons why we decided to use the rule-based system *FonPars* [12], and, e.g., not a memory-based learning system, e.g. [9]. FonPars is a rule compiler that can be used in combination with different rule sets. For a Dutch text-to-speech system, a standard Dutch rule set has been developed. If this (unadapted) standard Dutch rule set is applied directly to the enriched orthographic transcriptions of the current database, the resulting phone transcription will contain many errors (an example is given in table 4). Therefore, we will adapt this rule set for each interview or group of interviews with similar characteristics. How the adaptation should be carried out is one of the main research issues.

Transcription type	Transcription
Original enriched orthographic transcription	moar dan deden ze dat zo an banen ôver de dèle leggen
Phone transcription obtained with unadapted FonPars	mo-Ar dAn de-d@ z@ dAt so An ba-n@ ôvEr d@ dèl@ lE-g@
Corrected phone transcription	m@Ur dAn de-d@n z@ dAt so An ba-n@n ov@r d@ dE-l@ lE-g@n

Table 4. Different transcriptions of the example sentence from Aalten.

3.3. Word segmentation

On the basis of the phone transcription, a word segmentation will be derived in three steps:

1. Phone segmentation
2. Word segmentation
3. Post-processing

The phone segmentation will be generated with an HMM-based automatic speech recognition system. An important research issue constitutes which set of acoustic models should be used for segmentation. In this respect, it should be noted here that different sets of symbols are used in the CGN, FonPars, and the current dialect database. If different symbols are used for the same sounds, homogeneity can be obtained by replacing symbols were necessary. However, the symbols present in the CGN do not fully cover all the sounds present in (the transcriptions of) the dialects. Taking this into account, we will experiment with different sets of acoustic models, from broad to very detailed models: from acoustic models for classes (e.g., broad phonetic classes), to models for all symbols in the CGN. Another possibility would be to train acoustic models for dialect sounds that are not present in the CGN (like those in Table 2). However, this requires that training material is available in which these dialect sounds

occur frequently enough. Since it will be difficult to obtain a sufficient amount of training material, and since acoustic models for dialect sounds are probably not crucial for word segmentation, we will probably map the symbols of these dialect sounds on symbols of similar sounds present in the CGN. On the other hand, since the signal characteristics of the dialect database differ substantially from each other (mainly due to the fact that different sets of recording equipment were used), and those of the CGN, we will experiment with 'robustness' procedures that alleviate these differences in signal characteristics, such as pre-processing, acoustic adaptation, etc.

The result of the first step will be a phone segmentation, which will be converted to a word segmentation in the second step. In general, this means that only the first and last boundary of a word will be kept. In previous experiments on word segmentation, we have seen that systematic errors in the placement of boundaries occur (e.g., for plosives), and that, therefore, the quality of the resulting word segmentation can be improved by post-processing. Similar results were found by others [10]. Therefore, in the third step post-processing is carried out.

3.4. Quality control

The main goal of this project is to deliver a word segmentation. The phone transcription is only a byproduct used to obtain this word segmentation. During the development of the whole segmentation procedure, we will regularly take random samples in order to evaluate the quality of the phone transcriptions and the segmentations. At two times, i.e., when half and all of the data has been segmented, the automatically placed word boundaries will be compared with manually placed boundaries for about 5% of the data.

The two main types of errors that can occur are:

1. No segmentation can be calculated for a whole chunk.
2. A segment boundary is at the wrong place, i.e., the distance to the 'correct' place is above a threshold.

We will first focus on reducing the errors of type 1, and after that we will try to reduce the (average) distance between automatic and 'correct' boundary position. In the Dutch part of CGN, errors of type 1 occurred for about 20% of the chunks. This was especially the case when the chunks contained non-speech speaker sounds, such as laughing, coughing, etc. Although segmentation is probably more problematic for the material present in this dialect database, our goal is to reduce the amount of errors of type 1 to 20% or even lower. One of the methods we will explore, is to train acoustic models for non-speech speaker sounds.

4. Acknowledgements

The project is funded by the 'Digitaliseringsfonds' of the KNAW [17]. We would like to thank Loe Boves, Catia Cucchiarini, Diana Binnenpoorte, and Joop Kerkhoff for their useful contributions to this paper.

5. References

[1] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, M. Omologo, "Automatic Segmentation and Labeling of English and Italian Speech Databases". *Proceedings of Eurospeech 1993, pp. 653-656 Berlin, Germany.*

[2] V. Beattie S. Edmondson D. Miller Y. Patel and G. Talvola, "An Integrated Multi Dialect Speech Recognition System with Optional Speaker Adaptation", *Proceedings of Eurospeech 1999, pp. 1123-1126, Madrid, Spain.*

[3] N. Beringer, F. Schiel, "Independent Automatic Segmentation of Speech by Pronunciation Modeling", *Proceedings of the ICPhS 1999, pp. 1653-1656 San Francisco, USA..*

[4] L. Boves and N. Oostdijk, "Spontaneous Speech in the Spoken Dutch Corpus", *Proceedings ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR) 2003, Tokyo, Japan.*

[5] J. Brousseau and S. A. Fox, "Dialect-dependent speech recognisers for Canadian and European French", *Proceedings ICSLP 1992, pp. 1003-1006, Banff, Canada.*

[6] F. Brugnara, D. Falavigna and M. Omologo, "Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models", *Speech Communication, Vol. 12, pp. 357–370, 1993.*

[7] D. van Compernolle J. Smolders P. Jaspers and T. Hellemans, "Speaker Clustering for Dialectic Robustness in Speaker Independent Recognition" *Proceedings of Eurospeech 1991, pp. 723-726, Genova, Italy.*

[8] J. Daan and D.P. Blok, "Van randstad tot landrand", *Bijdragen en Mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam 37, 1969*

[9] W. Daelemans and A. Van den Bosch, "Language-independent data-oriented grapheme-to-phoneme conversion,", *In J. P. H. Van Santen, R. W.Sproat, J. P. Olive, and J. Hirschberg, (Eds.), pp. 77-89. Springer-Verlag, Berlin, 1996.*

[10] K. Demuynck, and T. Laureys, "A Comparison of Different Approaches to Automatic Speech Segmentation", *Proceedings of TSD 2002, pp. 277-284. Brno, Czech Republic.*

[11] T. Goeman and W. Jongenburger (in preparation), "Determinants of Fin de Siecle Dialect Use in the Netherlands", *International Journal of the Sociology of Language, 1997.*

[12] J. Kerkhoff, and T. Rietveld, "Prosody in Niros with Fonpars and Alfeios", In: de Haan and Oostdijk (Eds.), *Proceedings of the Dept. of Language & Speech, Univ. of Nijmegen, Vol.18, pp. 107-119, 1994.*

[13] H. C. Leung and V. W. Zue, "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech", *Proceedings of ICASSP 1984, pp. 2.7.1-2.7.4, San Diego, USA.*

[14] A. Ljolje and M. D. Riley, "Automatic segmentation and labeling of speech", *Proceedings ICASSP 1991, pp. S473-S476, Toronto, Canada.*

[15] J.P. Martens, D. Binnenpoorte, K. Demuynck, R. van Parys, T. Laureys, W. Goedertier & J. Duchateau, "Word Segmentation in the Spoken Dutch Corpus" *Proceedings of LREC2002, Las Palmas de Gran Canaria, Spain.*

[16] www.meertens.nl

[17] www.knaw.nl

[18] lands.let.kun.nl/projects/mumdid

[19] lands.let.kun.nl/cgn/ehome.htm

[20] www.praat.nl

[21] www.taal.phileon.nl/vergelijking.php