Title: Validation of phonetic transcriptions in the context of automatic speech recognition

Authors: Christophe Van Bael, Henk van den Heuvel, Helmer Strik

Affiliation: Centre for Language and Speech Technology
Radboud University Nijmegen, the Netherlands

Name and address for correspondence:
Christophe Van Bael
Radboud University Nijmegen
P.O. Box 9103
6500 HD Nijmegen
The Netherlands

Tel: + 31 24 361 29 08
Fax: + 31 24 361 29 07
E-mail: c.v.bael@gmail.com

Keywords: Broad Phonetic Transcriptions, Validation, Automatic Speech Recognition.

Abstract

Some of the speech databases and large spoken language corpora that have been collected during the last fifteen years have been (at least partly) annotated with a broad phonetic transcription. Such phonetic transcriptions are often validated in terms of their resemblance to a handcrafted reference transcription. However, there are at least two methodological issues questioning this validation method. Firstly, no reference transcription can fully represent the phonetic truth. This calls into question the status of such a transcription as a single reference for the quality of other phonetic transcriptions. Secondly, phonetic transcriptions are often generated to serve various purposes, none of which are considered when the transcriptions are compared to a reference transcription that was not made with the same purpose in mind. Since phonetic transcriptions are often used for the development of automatic speech recognition (ASR) systems, and since the relationship between ASR performance and a transcription's resemblance to a reference transcription does not seem to be straightforward, we verified whether phonetic transcriptions that are to be used for ASR development can be justifiably validated in terms of their similarity to a purpose-independent reference transcription.

To this end, we validated canonical representations and manually verified broad phonetic transcriptions of read speech and spontaneous telephone dialogues in terms of their resemblance to a handcrafted reference transcription on the one hand, and in terms of their suitability for ASR development on the other hand. Whereas the manually verified phonetic transcriptions resembled the reference transcription much closer than the canonical representations, the use of both transcription types yielded similar recognition results. The difference between the outcomes of the two validation methods has two implications. First, ASR developers can save themselves the effort of collecting expensive reference transcriptions in order to validate phonetic transcriptions of speech databases or spoken language corpora. Second, phonetic transcriptions should preferably be validated in terms of the application they will serve because a higher resemblance to a purpose-independent reference transcription is no guarantee for a transcription to be better suited for ASR development.

Abbreviations

ASR	Automatic Speech Recognition
CGN	Corpus Gesproken Nederlands – Spoken Dutch Corpus
MPT	Manual Phonetic Transcription
RT	Reference Transcription
WER	Word Error Rate

1. Introduction

Phonetic transcriptions are the written records of perceptual analyses of speech. They describe continuous speech signals as sequences of discrete phonetic symbols. These symbols can be chosen from small (more general) or large (more detailed) sets of symbols, depending on the purpose the transcriptions are generated for. Transcriptions can be handmade, machine-made or they can be generated through a joint effort of man and machine.

Formally speaking, the validity of phonetic transcriptions indicates the adequacy with which the transcriptions represent the original speech signals, and as such also the adequacy with which the transcriptions serve the purpose which they will be employed for (Cucchiarini, 1993). However, the purpose for which transcriptions are made is not always unique nor always known in advance. Some of the speech databases and large spoken language corpora that have been collected during the last fifteen years (e.g. Switchboard (Godfrey et al., 1992; Greenberg, 1997) or the Spoken Dutch Corpus (Oostdijk, 2002; Goddijn & Binnenpoorte, 2003)) have been (at least partly) annotated with a phonetic transcription without knowing the specific purpose(s) the transcriptions would serve, since the corpora were explicitly aimed at serving a wide variety of research and development projects. In such contexts, phonetic transcriptions can only be validated by means of a purpose-independent validation criterion.

More often than not, phonetic transcriptions are validated through a comparison with some handmade reference transcription (RT) that is considered to be the most accurate representation of the speech signal that can be obtained with a given set of transcription symbols. In the literature several different instantiations of RTs have been used. Saraçlar et al. (2000) used a manual transcription that was independently produced by a phonetician. Kipp et al. (1996) used several independently produced manual transcriptions, each of which served as an independent reference. Kuijpers and van Donselaar (1997) also used several independently produced manual transcriptions, but they used them as a single reference by considering only the majority vote for every phonetic symbol. Shriberg et al (1984) argued that the best possible transcription is obtained by forcing two or more expert phoneticians to agree on each and every symbol in the transcription. A so-called ‘consensus transcription’ differs from a majority vote transcription in that the latter does not involve a negotiation phase during which individual transcribers may change their original transcript. Irrespective of the procedure through which a reference transcription is obtained, we will call the validation of phonetic transcriptions in terms of their resemblance to an RT the *traditional validation method*.

There are at least two methodological issues that raise questions about the traditional method for validating phonetic transcriptions. The first issue relates to the status of the RT as the 'true' representation of the original speech signal. Since speech signals are the result of continuous dynamic gestures of articulators, each mapping of such a continuous process onto a sequence of symbols that are taken from a finite symbol set implies some degree of quantisation error. These errors show in the time domain as well as in the acoustic domain because all acoustic properties in a certain time interval have to be represented by just one symbol. Obviously, the quantisation errors in both domains will be larger if fewer symbols are used. The decision on the number and the identity of the symbols is to some extent dependent on the phonetician's background. It can be concluded that there is no such thing as the "true" representation of a speech signal in the form of a sequence of discrete symbols (Cucchiarini, 1993). Consequently, the concept of a unique symbolic representation of a speech signal is elusive at best. The traditional validation method, however, always requires such a unique representation in the form of a reference transcription.

The second methodological issue is less obvious. It is related to the seemingly undisputable operationalisation of the concept of a transcription's validity in terms of the transcription's similarity with a purpose-independent reference transcription; there may not always be such a clear correlation between a transcription's similarity to a reference transcription and the transcription's suitability to serve a certain purpose. For example, no matter what the accuracy of a broad phonetic transcription may be, it will not be suitable for a phonetician who wants to represent the degree of diphthongisation of long vowels, simply because a broad phonetic transcription only reflects two extreme stages of diphthongisation: the process is either fully present or completely absent. For other applications, in which the detail in the phonetic transcription seems to correspond to the detail required by the application, the usefulness of the traditional validation method may be more difficult to estimate in advance. One such application is the development of automatic speech recognition (ASR) systems.

ASR development requires large speech databases or spoken language corpora with corresponding phonetic transcriptions for several different purposes, including the training of acoustic models and the construction of pronunciation lexicons. It is intuitively reasonable to expect that acoustic models will be less polluted if they are trained on the basis of a 'better' transcription, and to think that words will be more accurately recognised if the recogniser's pronunciation lexicon comprises 'better' phonetic transcriptions. If we assume that transcriptions are 'better' if they are 'more similar' to a reference transcription, we assume that the traditional validation method is suitable for validating transcriptions that are to be used for ASR development.

Interestingly, however, the inverse relation between a transcription's resemblance to an RT and ASR performance does not hold. Kessens and Strik (2004) investigated the relationship between the performance of a set of continuous speech recognisers, and the resemblance between an RT and phonetic transcriptions that were generated by the different recognisers. They concluded that recognisers with a higher recognition performance (or a lower word error rate (WER)) do not guarantee the generation of phonetic transcriptions that are more similar to a given RT.

Since the relationship between recognition performance and a transcription's resemblance to an RT does not seem to be straightforward, this study was aimed at testing whether the traditional validation method offers a useful indication of a transcription's suitability for basic ASR development. If, in addition to the results in Kessens and Strik (2004), we would fail to find a positive relationship between a transcription's resemblance to an RT and its suitability to develop ASR systems, this would indicate that phonetic transcriptions may be better validated through an *application-oriented validation* method (which, in our case, would mean in terms of their contribution to ASR performance). Such a result would also indicate that ASR developers could save themselves the tedious and expensive effort of collecting reference transcriptions in order to validate phonetic transcriptions that may come with a new training database.

We required two resources to assess the validity of phonetic transcriptions in terms of their contribution to ASR performance. First, we required a corpus suitable for the training and the evaluation of an ASR system. This corpus had to contain at least two different transcriptions that could be used for that purpose. Second, we needed a fixed platform to develop and test the ASR system, in order to isolate the effect of the phonetic transcriptions from the multitude of other factors that could affect the performance of the ASR system.

Our first requirement was met by the Spoken Dutch Corpus (Oostdijk, 2002), a 9-million-word spoken language corpus, 10% of which comes with a manually verified broad phonetic transcription (Goddijn and Binnenpoorte, 2003). The second type of transcription that we used, viz. a canonical representation, is available in the canonical lexicon that typically comes with every corpus for ASR development. The corpus and the two transcriptions are described in more detail in Sections 3.1 and 3.2.

The requirement of a fixed platform to isolate the transcriptions as the only factor affecting the recognition performance was met by fixing the training and test corpora as well as the language models of our system. As a consequence, we could study the effect of the two transcription types in relation to 1) the amount of phonetically transcribed material that was used to train the acoustic models (since the production of manually verified transcriptions is time-consuming and expensive, the amount of training speech that comes with a manual phonetic transcription cannot be expected to be as large as the amount of speech that can be annotated with a canonical representation), 2) the procedures with which the acoustic models were trained (with the canonical representations, the manually verified phonetic transcriptions, or through a bootstrap procedure involving both transcription types), and 3) the pronunciations in the recognition lexicon (canonical representations or manually verified phonetic transcriptions).

Since we aimed at investigating the direct influence of the two transcriptions in a fixed experimental design, we did not aim at optimising recognition performance by all possible means. Rather, our intention behind the fixed experimental design was similar to the intention behind the research conducted in the framework of the AURORA project, where the ASR decoder was fixed, and performance improvements could only be obtained by adapting the acoustic features (Pearce, 2001). For the same reason, it should be clear that we did not aim at generating the most *accurate* transcription possible. Rather, we aimed at testing whether the traditional and the application-oriented validation method agreed on their assessments of the validity of the phonetic transcriptions in order to establish whether the traditional validation method guarantees an adequate indication of a transcription's suitability for ASR development.

This paper is organised as follows. Section 2 describes how canonical representations and manually verified phonetic transcriptions were validated in terms of the traditional validation method and in terms of their contribution to recognition performance. Section 3 presents the speech material and the architecture of the speech recogniser. In Section 4, we present and discuss the results of the validation experiments. In Section 5, we discuss the implications of our results.

2. Experimental setup

We validated canonical representations and manually verified phonetic transcriptions (MPTs) of data comprising two different speech styles: read speech and telephone dialogues. The details of the transcriptions are presented in Section 3.2. Here we confine ourselves to mentioning that the canonical representations were generated by concatenating the standard pronunciations of the words in the orthographic transcriptions. The MPTs were made by trained students who checked and corrected canonical representations by listening to the speech signal. The reference transcriptions were consensus transcriptions produced by two trained phoneticians.

2.1. THE TRADITIONAL VALIDATION METHOD

We compared the canonical representations and the manually verified phonetic transcriptions with reference transcriptions of the same data. To that end we aligned the transcriptions of every speech style with the appropriate RT. Subsequently we summarised the disagreements between the transcriptions and the RT in an overall disagreement measure that was defined as:

$$Percentage\ disagreement = \left(\frac{Sub_{phone} + Del_{phone} + Ins_{phone}}{N_{phone}} \right) \times 100\% \qquad (1)$$

i.e. the sum of all phone substitutions ($\mathrm{Sub}_{\mathrm{phone}}$), deletions ($\mathrm{Del}_{\mathrm{phone}}$) and insertions ($\mathrm{Ins}_{\mathrm{phone}}$) divided by the total number of phones in the RT ($\mathrm{N}_{\mathrm{phone}}$).

We used Align (Cucchiarini, 1996) to align the phonetic transcriptions and to compute the percentage disagreement between them. Align is a dynamic programming algorithm designed to compute the optimal alignment between two strings of phonetic symbols according to matrices in which the articulatory distances between the phonetic symbols are defined. The optimal feature matrices were determined in previous research on similar data (Binnenpoorte and Cucchiarini, 2003). The matrices are presented in Appendix 1.

2.2. THE APPLICATION-ORIENTED VALIDATION METHOD

We validated the canonical representations and the MPTs in terms of their contribution to the overall recognition performance of a standard continuous speech recogniser. We adhered to the traditional evaluation metric for recognition performance in ASR, the word error rate (WER), which is defined as:

$$WER = \left(\frac{Sub_{word} + Del_{word} + Ins_{word}}{N_{word}} \right) \times 100\% \qquad (2)$$

i.e. the sum of all word substitutions ($\mathrm{Sub}_{\mathrm{word}}$), deletions ($\mathrm{Del}_{\mathrm{word}}$) and insertions ($\mathrm{Ins}_{\mathrm{word}}$) divided by the total number of words in the orthographic reference transcription ($\mathrm{N}_{\mathrm{word}}$).

The overall recognition performance of a continuous speech recogniser can be influenced by numerous factors. Two important factors, viz. the quality of the acoustic models and the degree to which the pronunciation lexicon contains realistic phonetic transcriptions for words to be recognised, are directly dependent on the availability of suitable phonetic transcriptions. The quality of acoustic models depends on the suitability of the phonetic transcriptions of the training material, because acoustic model training involves a time-alignment of large amounts of speech with corresponding phonetic transcriptions. Likewise, the quality of a pronunciation lexicon is determined by the quality of its transcriptions, in that more realistic phonetic transcriptions increase the chance of words to be correctly recognised. In addition, it has repeatedly been found that recognition performance also depends on the (lack of) correspondence between the transcriptions in the recognition lexicon and the transcriptions with which the acoustic models are trained. As already indicated, we validated the canonical representations and the MPTs in terms of overall recognition performance. By fixing the continuous speech recogniser but for the acoustic models and the recognition lexicon, we guaranteed that differences in the overall recognition performance could only result from the transcriptions' influence on the acoustic models and the recognition lexicon.

Per speech style, we conducted a series of four experiments. In these experiments, we trained the same recogniser with different sets of acoustic models (all context-independent models with a fixed model topology, but trained with different transcriptions and different amounts of training data) and we tested the recogniser with different recognition lexica. Table 1 presents a schematic overview of the four experiments. The experiments were characterised by three variables: 1) the amount of training data we used to train the acoustic models (large or small training set), 2) the (combinations of) transcriptions we trained the acoustic models with (canonical, MPT or a bootstrap procedure involving both transcription types – see below) and 3) the type of the transcriptions in the recognition lexica (canonical or MPT).

	Size of the training sets	Transcriptions for the training of acoustic models	Transcriptions in the recognition lexica
Experiment 1	Small	Canonical	Canonical
Experiment 2	Small	MPT	MPT-based
Experiment 3	Large	Canonical	Canonical
Experiment 4a	Large	Bootstrap MPT + Canonical	Canonical
Experiment 4b			MPT-based

Table 1: *Overview of the recognition experiments.*

In experiment 1, we trained acoustic models with the canonical representations of the small training sets (see Section 3.1), and we used the same transcriptions to build canonical recognition lexica. The results of the first experiment formed a good baseline for the second experiment, in which we used the MPTs of the same small training sets to train the acoustic models and to build MPT-based recognition lexica. Since the production of MPTs tends to be time-consuming and expensive, larger sets of MPTs than the ones used in this second experiment are hardly ever available.

The third experiment resembled the first experiment, in that we trained acoustic models with canonical representations and in that we used the same canonical recognition lexica. However, this time we trained acoustic models with the canonical representations of much larger amounts of training data. The increased size of the data sets (as opposed to the first experiment) had to provide insight into the importance of the size of data sets for the training of efficient acoustic models. All acoustic models used in the first three experiments were generated from scratch (i.e. starting from a linear segmentation of the material).

In ASR, one often uses modest amounts of MPTs to train initial sets of acoustic models that, in a second training pass, are further trained with larger amounts of automatic phonetic transcriptions. This training method is called bootstrapping. We applied bootstrapping since we assumed that acoustic models that were initially trained with a small amount of MPTs and that were subsequently further trained with a large amount of canonical representations would outperform acoustic models that were trained from scratch with only canonical representations.

In the fourth experiment, we used the acoustic models of experiment 2 (which were trained on the MPTs of the small data sets) to align the speech data of the large data sets with the corresponding canonical representations of the data. Then we trained new acoustic models with the time-aligned canonical representations of the large data sets. Since the resulting acoustic models were based on a two-pass training procedure with MPTs and canonical representations, recognition experiments were carried out with both the canonical recognition lexica (exp. 4a) and the MPT-based lexica (exp. 4b). The alternating use of these recognition lexica (while using the same acoustic models) enabled us to study the effect of the different types of transcriptions in the recognition lexica in isolation.

To conclude, these experiments allowed us to validate the canonical representations and the manually verified phonetic transcriptions in terms of their suitability to train acoustic models and to generate recognition lexica. The transcriptions' suitability was reflected in and measured in terms of the recogniser's overall recognition performance. Whereas experiments 1 and 2 provided insight into the general influence of the two transcription types on the recognition performance, experiments 1 and 3 assessed the influence of different amounts of training data on the training of efficient acoustic models. Experiments 4a and 4b allowed us to investigate the influence of the different recognition lexica on the recognition performance.

3. Material and continuous speech recogniser

3.1 SPEECH MATERIAL

We extracted the speech material for our experiments from the Spoken Dutch Corpus (Corpus Gesproken Nederlands - CGN, 2004; Oostdijk, 2002). The Spoken Dutch Corpus is a 9-million-word multi-purpose spoken language corpus comprising Dutch as spoken in the Netherlands and Flanders in different communicative settings. The whole corpus was orthographically transcribed, lemmatised, and supplied with part-of-speech tagging. A 1-million-word subset of the corpus, the so-called *core corpus*, was enriched with a manually verified broad phonetic transcription and a syntactic annotation.

We conducted our experiments on speech from the Netherlands. The data comprised two speech styles with different acoustic and communicative properties: read speech (read aloud texts from a library for the blind) and conversational telephone dialogues. The read speech was recorded with table-mounted microphones and sampled at 16 kHz with a 16-bit resolution. The material comprised monologues with a vivid prosodic structure (due to the material's fictional content and the purpose the texts were read for: entertainment). The telephone dialogues were recorded through a telephone platform and sampled at 8 kHz with an 8-bit A-law coding. The two speakers in each conversation were recorded on separate channels.

		Reference sets	Experimental sets			
Speech style			Large training set	Small training set	Development test set	Evaluation test set
Read speech	# words	1,108	532,451	47,517	7,940	7,940
	hh:mm:ss	0:04:57	44:55:59	4:04:28	0:40:10	0:41:39
Telephone dialogues	# words	363	263,501	41,736	6,953	6,955
	hh:mm:ss	0:01:26	18:20:05	1:29:23	0:30:02	0:29:50

Table 2: *Statistics (number of words/tokens) of the data sets.*

Per speech style, we divided the material into two separate data sets which will hereafter be called the *reference sets* and the *experimental sets* (see Table 2). The data in the reference sets were provided with a consensus transcription. This enabled us to validate the phonetic transcriptions according to the traditional validation method. The data in the experimental sets were used to validate the phonetic transcriptions in terms of their suitability for ASR development (a more application-oriented validation method). To this end, the transcriptions were used to train (large and small training sets), tune (development test sets) and test (evaluation test sets) our continuous speech recogniser. Except for the training sets (the large training sets comprised the small training sets), all data sets were mutually exclusive.

3.2 PHONETIC TRANSCRIPTIONS

We worked with broad phonetic transcriptions of speech. All transcriptions were generated with the CGN phone set comprising 46 phones. However, not all of these phones occurred frequently enough in the training data to train robust acoustic models. In order to alleviate this problem, we mapped the phones in the transcriptions to the 39 phones presented in Appendix 2.

The canonical representations were generated by means of a lexicon-lookup procedure in which every word in the orthography was substituted with its standard pronunciation as represented in the canonical pronunciation lexica described in Section 3.3.1.

We extracted the MPTs of the data in the reference sets, the small training sets and the development and evaluation test sets from the CGN. The MPTs of the CGN are based on canonical representations to which all obligatory word-internal phonological processes (such as assimilation and degemination) were applied (Goddijn and Binnenpoorte et al., 2003; Booij, 1999). Cross-word processes were not applied. Human transcribers verified and corrected these example transcriptions according to a strict protocol. They were instructed to change the automatic transcriptions only if they were certain that the changes would yield a transcription that was substantially closer to the actual speech signal. As a consequence, the MPTs of the CGN may have a bias towards the canonical representations. However, such a check-and-correct procedure is a standard transcription procedure that has also been followed in other transcription projects (e.g. Greenberg, 1997).

The RTs were made in a fundamentally different way. Whereas the MPTs were made by human transcribers manually verifying an automatically generated transcription, the RTs were generated by two expert phoneticians transcribing from scratch. The transcribers had to reach a consensus on every symbol in the RTs. As a consequence, our reference sets were quite small compared to the evaluation test sets. However, whereas consensus transcriptions are always limited in size, they are often used to assess the validity of transcriptions obtained by means of other transcription procedures (like the MPTs and the canonical representations in our experiments).

3.3 LEXICA

3.3.1. Canonical pronunciation lexica

Our canonical lexica (one for each speech style) comprised one canonical pronunciation for every word in the development, evaluation and small training sets. The canonical lexica were compiled from the TST-lexicon (in-house version of 29-09-2004) and the CGN-lexicon. The TST-lexicon is a comprehensive multi-purpose lexicon for language and speech processing. It was compiled by merging various existing electronic lexical resources such as CELEX (Baayen et al, 1995), RBN (Referentiebestand Nederlands, 2005), and PAROLE (PAROLE lexicon, 2005). The CGN lexicon (delivered with the first release of the CGN) comprised the canonical representations of almost all unique word forms occurring in our data sets. The phonetic representations in the CGN lexicon were generated by means of TREETALK (Hoste et al., 2000), a grapheme-to-phoneme converter trained on the CELEX Dutch database (Baayen et al., 1995). Obvious errors in frequent words were manually corrected. The transcriptions of English loan words that were not yet included in the CGN lexicon were obtained from the CELEX English database (Baayen et al., 1995). The missing transcriptions of geographical names were obtained from ONOMASTICA (Quazza and van den Heuvel, 2000). The remaining out-of-vocabulary words were transcribed by means of a rule-based grapheme-to-phoneme converter (Kerkhoff and Rietveld, 1994) and the transcriptions were manually verified.

3.3.2. Pronunciation lexica with manually verified phonetic transcriptions

The MPT-based lexica (one for each speech style) were generated through word-to-transcription mappings between the orthographic transcriptions and the MPTs of the data in the development, evaluation and small training sets. We included the manually verified pronunciations of the words in the development and evaluation sets because not all of these words occurred in the small training sets. In doing so, we excluded the number of out of vocabulary words as an extra variable from the comparison of the canonical and the MPT-based lexica. Similarly, in order to exclude the lexical confusability from the comparison of the lexica, we retained only the most frequently observed pronunciation variant per word. This way both the canonical and the MPT-based lexica contained precisely one pronunciation for every word in the orthographic transcriptions.

The major difference between the canonical lexica and the MPT-based lexica was that the canonical lexica reflected the underlying morphological structure of the words and hypotheses about their underlying phonemic representations, whereas the MPT-based lexica mainly reflected knowledge about the most frequent pronunciation of the words in everyday speech. The MPT-based and the canonical lexica for the read speech contained different transcriptions for 40% of their entries, the lexica of the telephone dialogues for 45% of their entries.

3.4 THE CONTINUOUS SPEECH RECOGNISER

The continuous speech recogniser was built with the HTK toolkit (Young et al., 2001) using standard procedures. The characteristics of the recogniser were fixed in all experiments, except for the recognition lexicon and the acoustic models, which were based on the different phonetic transcriptions under investigation.

Several pre-processing procedures were applied to the speech signal. First pre-emphasis was applied. Feature extraction was implemented as a Fast Fourier Transform using a Hamming window every 10 ms for 25-ms frames. The mel-scaled filter bank analysis (50-8000 Hz for the read speech and 80-4000 Hz for the telephone dialogues) resulted in 39 cepstral coefficients per frame (12 coefficients and a separate energy component, and their delta and acceleration coefficients).

The recogniser used one back-off bigram language model per speech style. The evaluation test set perplexity of the read speech was 61.12. The evaluation test set perplexity of the telephone dialogues made 43.22. The lower test set perplexity of the telephone dialogues reflects the high frequency of standard phrases in the conversations. The higher test set perplexity of the read speech reflects the fact that the read speech comprised fragments with varied content from a number of different novels that were written by different authors. The order of magnitude of the test set perplexities was low enough to obtain credible WERs and at the same time high enough to not obscure the effects of improved acoustic models.

The acoustic models were 3-state continuous density left-right context-independent Hidden Markov Models. We trained speech style specific acoustic models on the canonical representations and the MPTs of the large and small training sets. Per set, 39 models were trained: 37 phone models, one model representing long silences, and one 1-state model modelling the optional short pauses between words (see Appendix 2). All models were gender-independent and accent-independent and comprised 32 mixture components (diagonal variance vectors) per state.

4. Results and discussion

4.1 TRADITIONAL VALIDATION METHOD

Table 3 reflects the validity of the phonetic transcriptions of both speech styles as assessed in terms of their overall disagreement (in % disagreement) with a reference transcription.

Speech style	PT	Substitutions (%)	Deletions (%)	Insertions (%)	% disagreement
Read speech	Canonical	7.39	3.51	1.14	12.04
	MPT	3.88	1.19	0.69	5.76
Telephone dialogues	Canonical	9.60	10.92	1.08	21.61
	MPT	4.68	2.64	1.08	8.4

Table 3: Validation of phonetic transcriptions in terms of their deviation from a reference transcription. The lower the disagreement, the better the transcription is considered to be.

The results in Table 3 are very clear: 1) the MPTs consistently resembled the RTs more than the canonical representations did ($p < .01$, *t*-test), and 2) the deviations of the different transcriptions from the RTs were larger when more spontaneous speech was involved. The significance of the differences suggests that the power of the test was sufficiently large despite the moderate size of the reference sets.

The relatively high resemblance between the MPTs and the RTs (as compared to the resemblance between the canonical representations and the RTs) is probably due to the fact that the MPTs and the RTs, even though produced according to different protocols (cf. Section 2.2), were produced by human transcribers who based their judgments on the actual speech signal. The canonical representations were automatically produced without taking the actual speech signal into account.

The results in Table 3 are in line with results published in the field. Binnenpoorte et al. (2003) also reported that the degree of resemblance between phonetic transcriptions and a reference transcription is inversely related to the degree of spontaneity of the transcribed speech, and proportional to the amount of manual effort devoted to the production of the transcriptions.

In any case, the results in Table 3 indicate that according to the traditional validation method, the validity of the MPTs of the Spoken Dutch Corpus is significantly higher than the validity of the canonical representations of the same material.

4.2. APPLICATION-ORIENTED VALIDATION METHOD

Table 4 reflects the validity of the phonetic transcriptions of both speech styles as assessed in terms of the transcriptions' contribution to recognition performance (in WER).

		Substitutions (%)	Deletions (%)	Insertions (%)	WER (%)
Experiment 1	Read speech	7.68	2.85	0.82	11.35
	Tel dialogues	33.43	17.12	2.60	53.16
Experiment 2	Read speech	7.95	2.07	1.27	11.28
	Tel dialogues	33.56	16.97	2.56	53.09
Experiment 3	Read speech	7.61	2.17	0.96	10.73
	Tel dialogues	32.47	17.97	2.13	52.57
Experiment 4a	Read speech	7.36	2.75	0.91	11.01
	Tel dialogues	33.64	16.99	2.66	53.30
Experiment 4b	Read speech	7.77	2.07	1.12	10.96
	Tel dialogues	33.26	17.11	2.52	52.42

Table 4: *Validation of phonetic transcriptions in terms of their influence on recognition performance. The lower the WER, the more suitable the transcription is considered to be.*

The modest nature of the recognition results in Table 4 can be partly explained by the lively prosody and fictional content characterising the read speech, and by the spontaneity and acoustic conditions characterising the telephone dialogues. Moreover, only bigram language models and context-independent acoustic models were used, since our main target, viz. validating phonetic transcriptions for ASR, only required the development of a standard recogniser that differed with respect to 1) the amount of phonetically transcribed data used to train the acoustic models, 2) the type of transcriptions of the training data, and 3) the type of transcriptions in the recognition lexicon. It is most striking that for both speech styles, none of the experiments yielded significantly different WERs ($p > .05$, *t*-test).

The recognition results of the first two experiments imply that the canonical representations were as suitable as the MPTs for training acoustic models on relatively small data sets (40K words), and for building pronunciation lexica for recognition. Remarkably, this did not only hold for the read speech, but also for the more spontaneous telephone dialogues in which the actual pronunciation could be expected to differ substantially from the canonical representation of the words. The MPT-based ASR system obtained a WER of 53.09%, which was almost identical to the 53.16% WER obtained by the system that was developed on the basis of the canonical representation of the words.

A comparison of the results of the first and the third experiment illustrates that the use of larger training sets (500K) decreased the WERs, though not significantly (0.62% absolute decrease on the read speech, 0.59% absolute decrease on the telephone dialogues). We did not conduct a similar experiment with MPTs, since the Spoken Dutch Corpus does not provide MPTs for such a large training set (nor does any other corpus available to date). However, MPTs of smaller data sets can be used to train acoustic models which in turn can be used to get good initial segmentations of much larger data sets. In our fourth experiment, we validated MPTs and canonical representations in terms of their potential for such a bootstrapping procedure.

In experiment 4a, we used the acoustic models trained on the MPTs of the small data sets (experiment 2) to get good initial segmentations of the large data sets. These segmentations were generated through a forced alignment of the canonical representations with the speech signal. A comparison of the results of experiments 3 and 4a illustrates that the bootstrapping procedure did not yield significantly different recognition results.

A comparison of the results of experiments 4a and 4b shows that the combined use of the MPT-based lexicon and the bootstrapped acoustic models yielded better (though not significantly better) results than the use of the canonical recognition lexicon with the same models. Especially the recognition of the telephone dialogues was facilitated by the use of the MPT-based lexicon. This is probably due to a larger mismatch between the actual data and the canonical representation of the spontaneous telephone speech.

At last, a comparison of the results of experiments 1 and 2 on the one hand and experiments 3, 4a and 4b on the other hand indicates that for both speech styles the acoustic models trained on the small data sets could not be improved substantially by adding more training material.

Overall, our recognition results are in line with a similar study on spontaneous telephone dialogues in American English (Switchboard) by Saraçlar et al. (2000). In that study, recognition experiments were conducted with different sets of acoustic models (trained on MPTs and automatic phonetic transcriptions) and matching decision tree-based pronunciation models. Their results showed that acoustic models trained on human transcriptions (Greenberg, 1997) did not give lower WERs than acoustic models trained on canonical baseforms. Saraçlar et al. (2000) found that the models trained on the MPTs gave lower phone error rates, but no lower WERs than the models trained on the canonical baseforms. They concluded that their results must have been due to the increased lexical confusability in the corresponding MPT-based recognition lexicon. Our results suggest that this cannot be the full explanation. By allowing only the most frequent transcription per word, we minimised the risk of increasing the lexical confusability. Still we observed similarly remarkable recognition results, which seem to suggest that for our ASR task, the canonical representations served their purpose as well as the manually verified phonetic transcriptions.

5. General discussion

This study was aimed at investigating whether the validity (or: the suitability) of phonetic transcriptions for basic ASR development can be assessed by means of the traditional validation method, i.e. in terms of the transcriptions' deviations from a handmade reference transcription. Previous research (Kessens and Strik, 2004) has shown that the relationship between recognition performance and a transcription's resemblance to an RT should not be taken for granted. In order to evaluate the usefulness of the traditional validation method, we conducted a series of experiments in which we assessed the influence of two different types of transcriptions (canonical representations and manually verified phonetic transcriptions) of two different speech styles (read speech and telephone dialogues) on the overall recognition accuracy of a continuous speech recogniser. As opposed to the traditional validation method, the assessment of the transcriptions' suitability for one particular purpose can be considered as an application-oriented validation method.

The outcome of the traditional validation method (which did not take into account the purpose the transcriptions would be used for) was quite outspoken: the validity of the MPTs was assessed much higher than the validity of the canonical representations because the MPTs deviated much less from the reference transcriptions than the canonical representations did. The application-oriented validation method gave quite another estimate of the transcriptions' validity. The assessment of the transcriptions' suitability for ASR showed that the use of MPTs and canonical representations did not yield significantly different recognition performance. This implies that both the MPTs and the canonical representations were equally valid for the purpose of developing a basic ASR system.

A comparison of the outcomes of the two validation methods supports different conclusions. First of all, it should be stressed that the application-oriented validation method did not contradict the usefulness of MPTs for ASR development, since we did not get better recognition results when using the canonical representations for this purpose. Logically, this also implies that the application-oriented validation method did not contradict the usefulness of manually verified transcriptions as such. As a matter of fact, for other purposes than training straightforward ASR systems (e.g. training more elaborate ASR systems), the story may well be different. For applications such as research in phonetics, it will probably even remain essential for transcriptions to reflect the speech signal as closely as possible. For such purposes, MPTs should definitely be preferred over canonical representations because canonical representations cannot (or only partially) represent the pronunciation variation observed in everyday speech.

A more important conclusion, however, is that the traditional validation method assigned a much higher validity rating to the MPTs than to the canonical representations. This was not confirmed by the outcome of our recognition experiment; the use of the canonical representations yielded similar recognition results. Considering the fact that the generation of MPTs is known to be time-consuming, expensive and error-prone (Cucchiarini, 1993), a preference for canonical representations seems more justified for our development task.

To conclude, we found no consistent relationship between the distance of a broad phonetic transcription to a reference transcription on the one hand, and the influence of that transcription on the recognition performance of a continuous speech recogniser on the other hand. This outcome has two implications. First of all, it suggests that ASR developers can save themselves the time and effort of collecting expensive reference transcriptions in order to validate phonetic transcriptions of speech databases or spoken language corpora. Second, and most importantly, it implies that phonetic transcriptions should preferably be validated in terms of the application they will serve because a higher resemblance to a purpose-independent reference transcription proved no guarantee for a transcription to be better suited for ASR development.

Acknowledgements

The work of Christophe Van Bael was funded by the Speech Technology Foundation (Stichting Spraaktechnologie, Utrecht, The Netherlands). The authors would like to thank Louis Pols, various colleagues at the Department of Language and Speech (now CLST) and three anonymous reviewers for their comments on previous versions of this paper.

References

1. Baayen, R.H., Piepenbrock, R., Gulikers, L. (1995). *The CELEX Lexical Database (Release 2).* Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
2. Binnenpoorte, D., Cucchiarini, C. (2003). Phonetic Transcription of Large Speech Corpora: How to Boost Efficiency without Affecting Quality. In: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 2981-2984.
3. Binnenpoorte, D., Goddijn S.M.A., Cucchiarini, C. (2003). How to Improve Human and Machine Transcriptions of Spontaneous Speech. In: *Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, Japan, pp. 147-150.
4. Booij, G. (1999). *The Phonology of Dutch.* Oxford University Press, New York.
5. CELEX Lexical Database (2005). [http://www.ru.nl/celex/].
6. Cucchiarini, C. (1993). *Phonetic Transcription: a Methodological and Empirical Study.* Ph.D. Dissertation, University of Nijmegen, the Netherlands.
7. Cucchiarini, C. (1996). Assessing Transcription Agreement: Methodological Aspects. In: *Clinical Linguistics and Phonetics*, vol. 10/2, pp. 131-155.
8. Goddijn, S.M.A., Binnenpoorte, D. (2003). Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus. In: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 1361-1364.
9. Godfrey, J., Holliman, E., McDaniel, J. (1992) SWITCHBOARD: Telephone Speech Corpus for Research and Development. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, USA, pp. 737-740.
10. Greenberg, S. (1997). The Switchboard Transcription Project. *Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series.* Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.
11. Hoste, V., Daelemans, W., Tjong Kim Sang, E., Gillis, S. (2000). Meta-learning for Phonemic Annotation of Corpora. In: *Proceedings of the 17th International Conference on Machine Learning (ICML)*, Stanford University, CA, USA, pp. 375-382.
12. Kerkhoff, J., Rietveld, T. (1994). Prosody in Niros with Fonpars and Alfeios. In: *Proceedings of the Department of Language and Speech, University of Nijmegen*, vol. 18, pp. 107-119.

13. Kessens, J.M., Strik, H. (2004). On Automatic Phonetic Transcription Quality: Lower Word Error Rates Do Not Guarantee Better Transcriptions. In: *Computer, Speech and Language*, vol. 18, pp. 123-141.

14. Kipp, A., Wesenick, M.-B., Schiel, F. (1996). Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP),* Philadelphia, USA, pp. 106-109.

15. Kuijpers, C., Donselaar, W. van. (1997). The Influence of Rhythmic Context on Schwa Epenthesis and Schwa Deletion in Dutch. In: *Language and Speech*, vol. 41/1, pp. 87-108.

16. Oostdijk, N. (2002). The Design of the Spoken Dutch Corpus. In: Peters, P., Collins, P., Smith, A. (Eds.) *New Frontiers of Corpus Research.* Rodopi, Amsterdam, pp. 105-112.

17. PAROLE lexicon. (2005). [http://ww2.tst.inl.nl].

18. Pearce, D. (2001). Developing the ETSI Aurora Advanced Distributed Speech Recognition Front-end & What Next? In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),* Madonna di Campiglio Trento, Italy, pp. 131-134.

19. Quazza, S., Heuvel, H., van den. (2000). Lexicon Development for Speech and Language Processing. In: Van Eynde, F., Gibbon, D. (eds.) *Lexicon Development for Speech and Language Processing.* Kluwer Academic Publishers, Dordrecht, pp. 207-233.

20. Referentiebestand Nederlands (RBN). (2005). [http://ww2.tst.inl.nl]

21. Saraçlar, M., Nock, H., Khudanpur, S. (2000). Pronunciation Modeling by Sharing Gaussian Densities across Phonetic Models. *Computer Speech and Language*, vol. 14, pp. 137-160.

22. Shriberg, L.D., Kwiatkowski, J., Hoffman, K. (1984). A Procedure for Phonetic Transcription by Consensus. In: *Journal of Speech and Hearing Research*, vol. 27, pp. 456-465.

23. Spoken Dutch Corpus - Het Project Corpus Gesproken Nederlands. (2005). [http://lands.let.kun.nl/cgn/ehome.htm].

24. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P. (2001). *The HTK book (for HTK version 3.1).* Cambridge University Engineering Department.

Appendix 1: Feature matrix used to align two phonetic transcriptions of speech (Align).

Consonant	Place	Voice	Nasal	Stop	Glide	Lateral	Fricative	Trill
p	5,0	1,0	0,0	0,5	0,0	0,0	0,0	0,0
b	5,0	2,0	0,0	0,5	0,0	0,0	0,0	0,0
t	4,0	1,0	0,0	0,5	0,0	0,0	0,0	0,0
d	4,0	2,0	0,0	0,5	0,0	0,0	0,0	0,0
k	2,0	1,0	0,0	0,5	0,0	0,0	0,0	0,0
f	5,0	1,0	0,0	0,0	0,0	0,0	0,5	0,0
v	5,0	2,0	0,0	0,0	0,0	0,0	0,5	0,0
s	4,0	1,0	0,0	0,0	0,0	0,0	0,5	0,0
z	4,0	2,0	0,0	0,0	0,0	0,0	0,5	0,0
x	2,0	1,0	0,0	0,0	0,0	0,0	0,5	0,0
G	2,0	2,0	0,0	0,0	0,0	0,0	0,5	0,0
m	5,0	2,0	0,5	0,0	0,0	0,0	0,0	0,0
n	4,0	2,0	0,5	0,0	0,0	0,0	0,0	0,0
N	2,0	2,0	0,5	0,0	0,0	0,0	0,0	0,0
l	4,0	2,0	0,0	0,0	0,0	0,5	0,0	0,0
r	3,0	2,0	0,0	0,0	0,0	0,0	0,0	0,5
w	5,0	2,0	0,0	0,0	0,5	0,0	0,0	0,0
j	3,0	2,0	0,0	0,0	0,5	0,0	0,0	0,0
h	1,0	2,0	0,0	0,0	0,0	0,0	0,5	0,0

Appendix 1a: *Articulatory distance between consonants.*

Vowel	Length	Place	Tongue	Round	Diphthong
i	1,5	3,0	4,0	1,0	1,0
I	1,0	2,5	3,5	1,0	1,0
e	2,0	3,0	3,0	1,0	1,5
@+	2,0	3,0	3,0	2,0	1,5
E	1,0	3,0	2,0	1,0	1,0
a	2,0	2,0	1,0	1,5	1,0
A	1,0	1,0	1,5	1,5	1,0
o	2,0	1,0	3,0	2,0	1,5
O	1,0	1,0	2,0	2,0	1,0
u	1,5	1,0	4,0	2,0	1,0
y	1,5	3,0	4,0	2,0	1,0
Y	1,0	2,5	3,5	2,0	1,0
@	1,0	2,0	2,5	1,5	1,0
E+	2,0	2,5	3,0	1,0	2,0
Y+	2,0	2,5	3,0	1,0	2,0
A+	2,0	1,5	3,0	2,0	2,0

Appendix 1b: *Articulatory distance between vowels.*

Appendix 2: Phone mapping 46 CGN phone set to 39 phone set.

Class	Example	CGN-symbol	Can/MPT symbol(s)
Plosives	put	p	p
	bad	b	b
	tak	t	t
	dak	d	d
	kat	k	k
	goal	g	k
Fricatives	fiets	f	f
	vat	v	v
	sap	s	s
	zat	z	z
	sjaal	S	S
	ravage	Z	z+j
	licht	x	x
	regen	G	G
	geheel	h	h
Sonorants	lang	N	N
	mat	m	m
	nat	n	n
	oranje	J	n+j
	lat	l	l
	rat	r	r
	wat	w	w
	jas	j	j
Short vowels	lip	I	I
	leg	E	E
	lat	A	A
	bom	O	O
	put	Y	Y
Long vowels	liep	i	i
	buur	y	y
	leeg	e	e
	deuk	2	@+
	laat	a	a
	boom	o	o
	boek	u	u
Schwa	gelijk	@	@
Diphthongs	wijs	E+	E+
	huis	Y+	Y+
	koud	A+	A+
Loan vowels	scène	E:	E
	freule	Y:	Y
	zone	O:	O
Nasalised vowels	vaccin	E~	E
	croissant	A~	A
	congé	O~	O
	parfum	Y~	Y
Long silence			sil
Optional short silence			sp