AUTOMATIC GENERATION OF PHONETIC TRANSCRIPTIONS FOR LARGE SPEECH CORPORA

Kris Demuynck, Tom Laureys

Katholieke Universiteit Leuven, ESAT-PSI
Kasteelpark Arenberg 10
3001 Leuven, Belgium
email: {Kris.Demuynck,Tom.Laureys}@esat.kuleuven.ac.be

Steven Gillis

University of Antwerp, GER-CNTS
Universiteitsplein 1
2610 Wilrijk, Belgium
email: Steven.Gillis@uia.ua.ac.be

ABSTRACT

We describe a method for the automatic production of phonetic transcriptions in large speech corpora. First, we focus on the application of different techniques for the generation of pronunciation variants. Then, we explain the application of a speech recognition system for selecting the acoustically best matching phonetic transcription. The system is evaluated on different test sets selected from the Spoken Dutch Corpus, ranging from read-aloud text to spontaneous speech, and achieves promising first results.

1. INTRODUCTION

In recent years a growing number of speech corpora has been developed for different languages. Work on English has revealed that these large corpora provide an indispensable information source for phonetic research [1], especially with respect to spontaneous speech phenomena, which rarely match text book knowledge. Yet, the manual phonetic annotation of corpora is a time-consuming, expensive and often tedious enterprise. As a result, manual annotation is mostly provided for only a small part of the corpus. This situation calls for automatic tools which provide annotations accurate enough to be used in large-scale phonetic research.

This paper describes the application of state-of-the-art automatic speech recognition (ASR) tools to the generation of phonetic transcriptions for Dutch. Initially, a full network of alternative phonetic transcriptions is generated for each sentence, based on orthographic information. This process draws on different techniques, from lexicon lookup to advanced grapheme-to-phoneme (g2p) conversion. Then, all phonetic alternatives are acoustically scored by a Viterbi pass through the speech recognizer and the most probable one is retained.

Our automatic phonetic transcriptions meet a double purpose. First, together with the speech signal they serve as input to an automatic speech segmentation algorithm. In the project Spoken Dutch Corpus an automatic word segmentation is provided for the complete database (10M words) [2]. At the same time, a manual phonetic transcription will be provided for a selection of only 1M words. This means that the automatic segmentation requires an automatic phonetic transcription procedure. Second, the inclusion of an automatic phonetic transcription for the complete corpus is considered within the Spoken Dutch Corpus project.

This publication was supported by the project "Spoken Dutch Corpus" (CGN-project) which is funded by the Flemish Government and the Netherlands Organization for Scientific Research (NWO).

⋆v	foreign words, without further indication of the language
⋆t	interjections (e.g. 'hum', 'pff', or 'aha')
⋆a	incomplete words, also with incomplete orthography
⋆x	ill-understood words (e.g. words deduced more from the context than from the acoustics)
⋆u	onomatopoeia (e.g. 'ring-ring') and mispronunciations
⋆z	Dutch words pronounced with a strong regional accent
⋆d	dialectal words
xxx	non-understood word(s)
ggg	well-audible or functional speaker sounds (e.g. coughing, laughing or screaming)

Table 1. The different markers used in the orthography of the Spoken Dutch Corpus.

Previous work on automatic phonetic transcriptions for Dutch based on ASR was reported in [3]. Yet, their focus lies on five specific phonological processes and on the problems involved in the production of reference transcriptions. This paper, on the other hand, describes and evaluates automatic phonetic transcriptions with respect to a single reference transcription. Frequently occurring deviations are reported and interpreted, while evaluation is performed on three test sets (49189 words in total) representing different degrees of spontaneity. In addition, all techniques employed to obtain the phonetic transcription are explained and evaluated.

2. AUTOMATIC PHONETIC TRANSCRIPTION OF CORPORA

2.1. Orthographic Input

The orthographic annotations in the Spoken Dutch Corpus are enriched with various markers providing specific pronunciation information. Table 1 shows the set of markers and their meaning. Next to these codes, 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.

2.2. Deriving Pronunciation Variants

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

Lexicon lookup: Fonilex [4] (200K entries) provides multiple phonetic 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 phonetic 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 missing proper nouns (currently $\pm$6000 entries), interjections, frequently used dialect words and items not covered in one of the other lexica.

Compounding, derivation and inflection: As Dutch is a morphologically productive language, lexica in itself are insufficient to cover all possible word forms: lexica are confined to the simple words (non-compounds) and the most frequent compounds and derivations. The pronunciation of all other Dutch words is found by decomposing the word into its basic constituents, concatenating the pronunciation of these constituents and applying a set of assimilation rules. The decomposition starts with the simplest rule (a two-word compound) and activates more complicated rules (up to a three-word compound with inflection) until at least one decomposition is found. In our approach all decompositions possible based on pure orthographic constraints are pursued, i.e. no syntactical constraints are imposed. So some degree of overgeneration is introduced (e.g. 'varkensteelt' $\rightarrow$ 'varkens' + 'teelt' / 'varken' + 'steelt'). This overgeneration however rarely resulted in new pronunciation variants and even showed to be useful for handling Dutch proper nouns and mispronunciations.

Abbreviations and digits: At the moment, abbreviations are phonetically transcribed as the concatenation of the constituent letter word transcriptions. Exceptions (e.g. NATO) are currently added to one of the specific lexica, but will be processed in a more intelligent way in the future. The pronunciation of numbers inside the abbreviations is solved with a rule-based system.

Broken-off words: Broken-off words 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 dialectical 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 [5] and trained on the Fonilex database. More information on the configuration of the g2p system will be given in section 4.2.

Assimilation: The aforementioned resources provide phonetic transcriptions for all words in the corpus, in the case of Fonilex this is an abstract phonetic transcription which reflects multiple plausible transcriptions. In continuous speech however, phonemes at word ends have an influence on 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 [6], resulting in a compact pronunciation network. The set of rules used are a subset of the rules defined in the Fonilex database (most word-internal assimilation rules also operate across word boundaries), extended with rules found in other resources [7].

As to the application frequency of the above resources: in the case of the Spoken Dutch Corpus, at least for the Flemish data released up till now (1.8 million words), 17% of the words in the word list (70562 entries) are new compounds and inflections derived from the Fonilex lexicon. The remaining words not covered by Fonilex account for another 17%, distributed as follows: proper nouns (8.2%), foreign words (2.8%), incomplete words (2.3%), new words, dialect words, onomatopoeia and mispronunciations (2.3%), and abbreviations (1.3%).

2.3. Selecting the Best Matching Pronunciation

Once a pronunciation network is generated for every sentence, the transcription matching best with the speech signal must be selected automatically. All phonetic alternatives are acoustically scored (maximum likelihood) in a single pass (Viterbi alignment) through our speech recognition system using context-independent or 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 [6]. Details on the acoustic models will be given in section 4.3.

3. THE MANUAL REFERENCE TRANSCRIPTION

The automatic phonetic transcriptions were evaluated on a comparison with a single manual reference transcription. This section describes how the manual transcription was produced in the Spoken Dutch Corpus project. The general aim was to obtain a broad representation of the speech signal using a (slightly adapted) SAMPA notation [8]. A two-step procedure was developed: (1) a broad phonetic transcription was automatically derived from the orthographic transcription, and (2) this transcription was manually verified and corrected using the speech signal.

3.1. Automatic Transcription

The selected orthographically transcribed material was automatically transcribed phonetically. For this purpose a classifier was trained using the Fonilex database. The classifier incorporates the k-nn algorithm with information gain ratio feature weighting as implemented in the TiMBL software package [9]. Detailed information concerning the phoneme classifier can be found in [10]. Cross-word phonological phenomena were not covered.

3.2. Manual Verification

For the actual verification process, the automatic transcriptions were used together with the speech signal. The aim was to verify and to correct the transcriptions if necessary. The manual verification of the transcribed speech was organized in two steps: in a first phase the transcriptions were verified by a research assistant (RA) and in a second phase they were checked again by a project collaborator supervising the complete transcription cycle.

The RAs were recruited among linguistics students enrolled in a phonetics class, and hence had a training in narrow phonetic

transcription. They took part in a one hour instruction session during which the transcription manual was scrutinized and the transcription tool (viz. Praat) was demonstrated. A short standardized transcription test was then administered in order to check the assistant's proficiency.

Given this procedure in which every fragment is verified twice, a time investment of factor 26 is required, i.e. an hour of speech requires 26 hours verification time, distributed over 18 hours for the RAs and an additional 8 hours for the project coordinator. These figures vary according to the type of speech: the real-time factor (number of hours required for an hour of speech) ranges from factor 15 for formal lectures and speeches to factor 40 for spontaneous face-to-face dialogues and the like.

4. EXPERIMENTS

4.1. Test Sets

Experiments were performed on data taken from the Spoken Dutch Corpus. Three test sets were selected, representing different degrees of difficulty. Test set 1 accounts for the cleanest speech in the corpus, namely the read-aloud texts. We selected 50 speakers for this set, resulting in 1 hour and 37 minutes of speech (14176 words). Broadcast material (documentaries, news shows, ...) and public speeches belong to test set 2 (45 minutes, 7135 words) which includes 23 different speakers. The fragments in test set 2 are generally harder to process as the speaker's style becomes more disfluent and background noise might be present. Finally, test set 3 consists of informal interviews, discussions and school lessons (2 hours and 53 minutes, 27878 words) partaken in by 11 different speakers. This test set poses the hardest problem for a system producing automatic transcriptions as it is riddled with dialectical pronunciations, incomplete words, overlapping speech, etc.

4.2. Grapheme-to-Phoneme Conversion

As mentioned earlier, an induction decision tree (ID3) system was used when extended lookup failed to map the grapheme sequence to a phoneme sequence. Each phoneme is predicted based on a vector of 10 variables: the grapheme under consideration, a context of four left and four right graphemes and the last decoded phoneme (feedback). Larger contexts did not improve the results any further. Phonetic transcriptions are generated from back to front so that the last decoded phoneme corresponds to the right neighbour which turned out to be most informative.

The alignment between graphemes and phonemes in the training database (Fonilex) is performed by means of dynamic programming which allows for an arbitrary number of deletions and up to two insertions in a row. The cost functions, i.e. probability functions for grapheme-phoneme correspondences and conditional insertion probability functions, were trained in a maximum likelihood fashion by looping through 10 iterations of alignment and re-estimation. To be able to cope with unseen events, the cost functions were smoothed (1) based on distance metrics between the phonemic classes for substitution and deletion phenomena, and (2) based on empirically derived phonemic compatibility metrics for insertion phenomena.

The first version of the g2p system, which used the individual letters in the words as graphemic units, obtained an error rate of 6.5% on the word level for a 10-fold cross-validation experiment on the Fonilex database. The second and final version of the system clusters typical letter sequences (see table 2) depending on the orthographic context in order to obtain smarter graphemic units. This lowers the error rate to 6.0%.

pp bb tt dd kk cc gg ss zz ff mm nn ll rr
ng qu gh gn ck ch sh th
aa oo uu ee ie eu oe ui ei ij ou au
uw ow aw ay oy ey ai oi io oa ae ea ue
ouw auw oeu eau ooi aai oei
euill euil ueil

Table 2. Combining letters into clusters improves the grapheme-to-phoneme converter. Note that the selection of the clusters is context-sensitive, so that for example 'geuit' is split as 'g e ui t' instead of 'g eu i t'.

model	test set 1	test set 2	test set 3
unconstrained phoneme recognition			
CI	32.79%	40.62%	54.83%
CD	29.25%	36.96%	51.28%
constrained phoneme recognition			
CI	5.56%	6.54%	8.78%
CD	4.99%	6.15%	8.25%

Table 3. Results for unconstrained and constrained phoneme recognition.

4.3. Constrained Phoneme Recognition

Different acoustic models were evaluated based on the accuracy of the produced phonetic transcriptions. The acoustic models were trained on 7 hours of read newspaper text taken from the CoGeN database [11]. First, 3-state left-to-right context-independent (CI) phoneme models were created. The input features consist of 13 Mel-warped and mean-normalized cepstral coefficients with their first and second order time derivatives. The density functions for the 142 states (46 3-state phoneme models and 4 single-state models for noise, filled noise, speaker noise and garbage) are gaussian mixture models with 600 components on average. However, since most gaussians are shared between the different phoneme models [6], the total number of gaussians is limited to 10241. Next, context-dependent (CD) models were created based on these context-independent models using a decision tree approach. This results in models with 667 states and mixtures models with 115 components on average. The total number of gaussians is unchanged with respect to the context-independent models.

As a reference, the accuracy of both models was evaluated by means of an unconstrained phoneme recognition experiment, i.e. no information whatsoever concerning the orthography was used. The only additional information source used besides the acoustic models was a bigram phoneme transition model. The accuracy (sum of insertions, deletions and substitutions) obtained by both models is given in the top part of table 3. In total, 58026, 30163 and 99882 phonemes needed to be recognized for test set 1, 2 and 3 respectively. Next, both models were used to select from the pronunciation network the single phoneme sequence matching best with the speech signal (constrained phoneme recognition). The result of these experiments are given in the bottom part of table 3.

Both experiments show a clear gradation as to the level of dif-

type	freq.	details & relative importance	
ins.	0.78%	n: 22.7% ə: 19.0% j/w: 10.3% l/r: 8.7% t/d: 7.9%	
del.	1.28%	h: 25.1% ə: 23.6% n: 22.5% t/d: 10.0% j: 8.1%	
subst.	2.94%	inter vowel:	45.8%
		long → short vowels:	16.6%
		short → long vowels:	5.5%
		ə & əː ↔ ɛː, e, E, I:	10.8%
		inter consonant:	49.2%
		unvoiced → voiced:	17.9%
		voiced → unvoiced:	14.9%
		nasals (n, m, ŋ):	6.8%

Table 4. Detailed analysis of the most frequent errors for test set 1.

ficulty of the three test sets. Note however that part of the degradation is not due to the increase in difficulty but due to increasing mismatch between training conditions (read speech) and test conditions (spontaneous speech). The results also show that context-dependent models are to be preferred. This is not as trivial as it may sound, since context-dependent models have the inherent ability to model co-articulation effects such as sound assimilation or insertions, and may thus no longer correspond uniquely to the specific phoneme they are supposed to model.

Table 4 gives an analysis of the most frequent insertion, deletion and substitution phenomena in test set 1. The other test sets show very similar patterns. A detailed study of the contexts in which these insertion and deletion phenomena occur showed that not every deviation was a mistake on the side of the automatic system. Humans tend to hear what they expect and only by scrutinizing the acoustic signal subtle phonetic phenomena can be observed. Moreover, the Spoken Dutch Corpus is a large corpus and the human transcribers must work at a considerable speed, so they tend to focus on the more frequent phenomena. We list some of the less common phenomena that were frequently overlooked by human transcribers but were detected by the automatic procedure:

- ə-insertion in non-homorganic consonant clusters in coda position [12] (‘scherp’ /sxɛrəp/).
- Homorganic glide insertions between vowels (‘die een’ /dijən/).
- ə-deletion of the first ə in two consecutive syllables headed by a ə, provided that the resulting consonant cluster is a plosive followed by a liquid [12] (‘latere’ /latrə/).
- n-deletion due to nasal assimilation (‘onmacht’ /ɔmɑxt/).

Other errors are due the acoustic models or missing assimilation rules. Some of the problems with the acoustic models are (1) the minimal duration constraint of the 3-state left-to-right models (30 msec), (2) the mismatch between the train and test conditions, and (3) the fact that the context-dependent models are somewhat contaminated since no assimilation effects were taken into account in the training process.

5. CONCLUSIONS

We described a method for the automatic generation of phonetic transcriptions in large speech corpora. The system first applies different techniques for the generation of all plausible pronunciation variants for a given orthographic transcription. Next, a speech recognition system is employed for selecting the acoustically best matching transcription. Evaluation of the errors made by this system shows that the automatic system was not always to blame: humans also make mistakes, e.g. due to tiredness and loss of concentration, phenomena which never trouble automatic systems. To limit the error count on behalf of the automatic system, the following actions can (and will) be taken: (1) retraining of the models as to eliminate the mismatch between train and test conditions and to obtain cleaner acoustic models by applying the assimilation rules during the training phase as well, (2) the introduction of 2-state or single state models for phonemes that tend to be pronounced very rapidly, and (3) the derivation of more assimilation rules based on what is observed in the Spoken Dutch Corpus. But even without these modifications, the results obtained by the automatic system are up to standard.

6. REFERENCES

[1] P.A. Keating, “Word-level phonetic variation in large speech corpora,” in *Proc. of ‘The Word as a Phonetic Unit’*, A. et al. Alexiadou, Ed., number 11 in ZAS Papers in Linguistics, pp. 35–50. 1998.

[2] N. Oostdijk, “The Spoken Dutch Corpus,” *The ELRA Newsletter*, vol. 5, no. 2, pp. 4–8, 2000, Available from http://lands.let.kun.nl/cgn/home.htm.

[3] M. Wester, J.M. Kessens, C. Cucchiarini, and H. Strik, “Obtaining phonetic transcriptions: A comparison between expert listeners and a continuous speech recognizer,” *Language and Speech*, vol. 44, no. 3, pp. 377–403, 2001.

[4] P. Mertens and Filip Vercammen, “FONILEX manual,” Technical report, K.U.Leuven – CCL, 1998, http://bach.arts.kuleuven.ac.be/fonilex/.

[5] V. Pagel, K. Lenzo, and A.W. Black, “Letter to sound rules for accented lexicon compression,” in *Proc. ICSLP*, Sydney, Australia, 1998, vol. I, pp. 252–255.

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

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

[8] S. Gillis, C. Cucchiarini, S. Goddijn, and L. Pols, “Protocol voor brede fonetische transcriptie.,” Cgn document, 2001, http://www.elis.rug.ac.be/cgn/.

[9] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch, “TiMBL: Tilburg memory based learner, version 4.0, reference guide,” ILK Technical Report 01-04, Tilburg University, 2001.

[10] V. Hoste, W. Daelemans, and S. Gillis, “A rule induction approach to modeling regional pronunciation variation,” in *Proceedings of COLING*. 2000, pp. 327–333, San Francisco: Morgan Kaufman Publishers.

[11] K. Demuynck, D. Van Compernolle, C. Van Hove, and J.-P. Martens, “CoGeN een corpus gesproken nederlands voor spraaktechnologisch onderzoek — eindverslag,” Tech. Rep., K.U.Leuven - ESAT & Universiteit Gent - ELIS, 1997.

[12] G.E. Booij, *The Phonology of Dutch*, Clarendon Press, Oxford, 1995.