The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database

R.J.J.H. van Son[1], Diana Binnenpoorte[2], Henk van den Heuvel[2], and Louis C.W. Pols[1]

[1]Institute of Phonetic Sciences (IFA) / ACLC, University of Amsterdam, the Netherlands
[2]SPEX/A2RT, Nijmegen University, the Netherlands
Rob.van.Son@hum.uva.nl

Abstract

An open source database of hand-segmented Dutch speech was constructed with off-the-shelf software using speech from 8 speakers in a variety of speaking styles. For a total of 50,000 words, speech acquisition and preparation took around 3 person-weeks per speaker. Hand segmentation took 1,000 hours of labeling altogether. The asymptotic segmentation speed was about one word, or four boundaries, per minute. An evaluation showed that the *Median Absolute Difference* of the segment boundaries was 6 ms between labelers, and 4 ms within labelers. Label differences (substitutions, insertions, and deletions) were found in 8% of the segments between labelers and 5% within labelers. Compiled data are available in relational database format for querying with SQL.

1. Introduction

More and more large speech databases are becoming available for speech research and commercial R&D ([6], e.g., [3], [5], [10], [12], [13], [15]). However, the speech corpora currently available (e.g., Switchboard, Speechdat, RM) typically are collected through telephone networks ([5], [6]), have only a limited number of styles, use many speakers only once, and are not segmented at phoneme level (c.f., [5], [6], [10]). Furthermore, they tend to be expensive. What is typically needed for phonetic research is: phonemic (or phonetic) transcription and segmentation, broadband recording, and a lot of speech from each speaker. Also, (re-)distribution should be free. Currently, for Dutch a few speech corpora exist which more or less approximate these requirements: the Groningen corpus [6], EUROM [16], and the Spoken Dutch Corpus (CGN) [12], [13]. However, the first two have only limited speech styles and the latter is not ready yet. None of these corpora have phonemic segmentation, nor are the same speakers recorded in many styles. This dearth of segmented corpora for Dutch can be replicated for almost any other language. Whether or not segmented speech corpora are generally available depends on personal initiatives of individual researchers (c.f., [3], [15]).

One of the reasons hand-segmented speech corpora are lacking is the perceived costs of creating them. These costs are almost completely determined by the segmentation effort. For a limited number of speakers, the cost of recording informal and read speech in the laboratory is not prohibitive. Text preparations, recording, orthographic transliteration, automatic phonemic transcription, and an automatic alignment with a "standard" HMM speech recognizer can all be handled in less than 3 person-weeks per speaker involving about 90 minutes mixed style speech per speaker. However, an expensive "hand-correction" of the segmentation is needed before a corpus can be used for phonetics research. The received opinion is that the hand-alignment of phonemes costs (much) more than the preceding factors combined. In this paper we would like to introduce the IFA corpus and present some experience-based facts about the costs and benefits of hand-segmented corpora to help making informed decisions.

2. Corpus purpose

In the context of a phonetics project on the factors influencing intra-speaker variation of speech we had a need for a labeled and segmented corpus with broadband Dutch speech, with speech in a variety of styles (e.g., informal, read, isolated words). It was decided to construct a "reusable", general purpose, 50,000 word corpus. This was seen as a good opportunity to study the real costs and trade-offs involved in the construction of a corpus of hand-segmented speech to benefit future projects (e.g., the INTAS project [4], [13]).

Access and distribution of the available large databases are quickly becoming a problem. For instance, the complete Spoken Dutch Corpus (*CGN* [12], [13]), containing a wide range of speaking styles and speakers, will, for the time being, be distributed on about 175 CD-ROMs, making on-site management a real challenge. The history of database projects in the sciences (e.g., biology) shows that most users treat these corpora as "on-line libraries" where they look for specific information (c.f., [2]). Most queries are directed towards compiled data, not towards raw data. Many journals (e.g., Nature [9]) also require that raw and compiled data underlying publications be made available through a publicly accessible database. We can expect developments in a similar direction in speech and language research.

From the experiences in the sciences, some general principles for the construction and management of large corpora can be distilled that were taken as the foundation of the architecture of the IFA corpus:

- Access should be possible using a powerful query language [2], [3]
- Basic data should be available in compiled form
- Internet access is indispensable
- "Reviewed" user contributions should be stimulated and incorporated

3. Corpus construction

3.1. Speakers

Speakers were selected at the Institute of Phonetic Sciences in Amsterdam (IFA) and consisted mostly of staff and students. Non-staff speakers were paid. In total 18 speakers (9 male, 9 female) completed both recording sessions. All speakers were mother-tongue speakers and none reported speaking or hearing problems. Recordings of 4 women and 4 men were

selected for phonemic segmentation, based on distribution of sex and age, and the quality of the recordings. The ages of the selected speakers ranges from 15 to 66 years of age (Table 1).

Table 1: Corpus contents (excluding empty and filled pauses). Printed are the number of items. The segmented items are a subset of the recorded items. S: Sentences and sentence-sized collections, W: Words, Sy: Syllables, Ph: Phonemes.

Speaker	Recorded		Segmented			
sex/age	S	W	S	W	Sy	Ph
N F/20	1078	11013	727	7644	11108	28043
G F/28	832	10944	806	10315	14683	36807
L F/40	640	8753	542	6882	10087	25344
E F/60	873	11246	712	8654	12896	32715
R M/15	655	7106	453	4621	6560	16015
K M/40	602	7667	400	4610	6577	15971
H M/56	675	8101	536	6444	9039	23190
O M/66	773	8237	316	2612	3752	9459
all	6128	73067	4492	51782	74702	187544

Each speaker filled in a form with information on personal data (sex, age), socio-linguistic background (e.g., place of birth, primary school, secondary school), socio-economic background (occupation and education of parents), physiological data (weight/height, smoking, alcohol consumption, medication), and data about relevant experience and training.

3.2. Speaking styles

Eight speaking "styles" were recorded from each speaker (Table 2). From informal to formal these were:

1. Informal story telling face-to-face to an "interviewer" (**I**)
2. Retelling a previously read narrative story without sight contact (**R**)

And reading aloud:

3. A narrative story (**T**)
4. A random list of all sentences of the narrative stories (**S**)
5. "Pseudo-sentences" constructed by replacing all words in a sentence with randomly selected words from the text with the same POS tag (**PS**)
6. Lists of selected words from the texts (**W**)
7. Lists of all distinct syllables from the word lists (**Sy**)
8. A collection of idiomatic (the Alphabet, the numbers 0-12) and "diagnostic" sequences (isolated vowels, /hVd/ and /VCV/ lists) (**Pr**)

The last style was presented in a fixed order, all other lists (S, PS, W, Sy) were (pseudo-)randomized for each speaker before presentation.

Each speaker read aloud from two separate text collections based on narrative texts. During the first recording session, each speaker read from the same two texts (*Fixed* text type). These texts were based on the Dutch version of "The north wind and the sun" [14], and on a translation of the fairy tale "Jorinde und Joringel" [8]. During the second session, each speaker read from texts based on the informal story told during the first recording session (*Variable* text type). A non-overlapping selection of words was made from each text type (W). Words were selected to maximize coverage of phonemes and diphones and also included the 50 most frequent words from the texts. The word lists were automatically transcribed into phonemes using a simple CELEX [17] word list lookup and were split into syllables. The syllables were transcribed back into a pseudo-orthography which was readable for Dutch subjects (Sy). The 70 "pseudo-sentences" (PS) were based on the *Fixed* texts and corrected for syntactic number and gender. They were "semantically unpredictable" and only marginally grammatical.

Table 2: Distribution of segmented words per speaker over speaking styles (I-Pr, see text). Silent and filled pauses are excluded. Last two rows show the corresponding mean articulation rate per sentence in syllables/s (Sy) and phonemes/s (Ph).

Sp	I	R	T	S	PS	W	Sy	Pr
N	660	385	2427	2850	412	262	292	356
G	1850	1639	2761	2868	206	230	290	470
L	885	465	2126	2078	423	239	274	387
E	933	1178	2556	2765	215	261	313	432
R	127	323	1348	1449	451	232	268	423
K	538	435	1354	1346	-	248	275	415
H	269	658	2005	2081	435	259	286	451
O	-	1173	-	-	466	253	284	436
all	5262	6256	14577	15437	2608	1984	2282	3370
Sy	*5.5*	*5.2*	*5.7*	*5.6*	*4.6*	*3.5*	*2.4*	*3.5*
Ph	*13.5*	*13.1*	*14.4*	*14.3*	*12.2*	*9.3*	*6.7*	*6.3*

3.3. Recording equipment and procedure

Speech was recorded in a quiet, sound treated room. Recording equipment and a cueing computer were in a separated control room. Two-channel recordings were made with a head-mounted dynamic microphone (Shure SM10A) on one channel and a fixed HF condenser microphone (Sennheiser MKH 105) on the other. Recording was done directly to a Philips Audio CD-recorder, i.e., 16 bit linear coding at 44.1 kHz stereo. A standard sound source (white noise and pure 400 Hz tone) of 78 dB was recorded from a fixed position relative to the fixed microphone to be able to mark the recording level. The head mounted microphone did not allow precise repositioning between sessions, and was even known to move during the sessions (which was noted).

On registration, speakers were given a sheet with instructions and the text of the two fixed stories. They were asked to prepare the texts for reading aloud. On the first recording session, they were seated facing an "interviewer" (at approximately one meter distance). The interviewer explained the procedure, verified personal information from a response sheet and asked the subject to tell about a vacation trip (style I). After that, the subject was seated in front of a sound-treated computer screen (the computer itself was in the control room). Reading materials were displayed in large font sizes on the screen.

After the first session, the subject was asked to divide into sentences and paragraphs a verbal transcript of the informal story told. Hesitations, repetitions, incomplete words, and filled pauses had been removed from the verbal transcript to allow fluent reading aloud. No attempts were made to "correct" the grammar of the text. Before the second session, the subject was asked to prepare the text for reading aloud. In the second session, the subject read the transcript of the informal story, told in the first session.

The order of recording was: Face-to-face story-telling (I, first session), idiomatic and diagnostic text (Pr, read twice), full texts in paragraph sized chunks (T), isolated sentences (S), isolated pseudo-sentences (PS, second session), words (W)

and syllables (Sy) in blocks of ten, and finally, re-telling of the texts read before (R).

3.4. Speech preparation, file formats, and compatibility

The corpus discussed in this paper is constructed according to the recommendations of [6], [7]. Future releases will conform to the *Open Languages Archives* [1]. Speech recordings were transferred directly from CD-audio to computer hard-disks and divided into "chunks" that correspond to full cueing screen reading texts where this was practical (I, T, Pr) or complete "style recordings" where divisions would be impractical (S, PS, W, Sy, R).
Each paragraph-sized audio-file was written out in orthographic form conform to [7]. Foreign words, variant and unfinished pronunciations were all marked. Clitics and filled pause sounds were transcribed in their reduced orthographic form (e.g., *'t, 'n, d'r, uh*). A phonemic transcription was made by a lookup from a CELEX word list, the pronunciation lexicon. Unknown words were hand-transcribed and added to the list. In case of ambiguity, the most normative transcription was chosen.
The chunks were further divided by hand into sentence-sized single channel files for segmenting and labeling (16 bit linear, 44.1 kHz, single-channel). These sentence-sized files contained real sentences from the text and sentence readings and the corresponding parts of the informal story telling. The retold stories were divided into sentences (preferably on pauses and clear intonational breaks, but also on "syntax"). False starts of sentences were split off as separate sentences. Word and syllable lists were divided, corresponding to a single cueing screen of text. The practice text was divided corresponding to lines of text (except for the alphabet, which was taken as an integral piece). Files with analyses of pitch, intensity, formants, and first spectral moment (center of gravity) are also available.
Audio recordings are available in AIFC format (16 bit linear, 44.1 kHz sample rate), longer pieces are also available in a compressed format (Ogg Vorbis). The segmentation results are stored in the (ASCII) label-file format of the *Praat* program (http://www.praat.org).
Label files are organized around hierarchically nested descriptive levels: phonemes, demi-syllables, syllables, words, sentences, paragraphs. Each level consists of one or more synchronized *tiers* that store the actual annotations (e.g., lexical words, phonemic transcriptions). The system allows an unlimited number of synchronized tiers from external files to be integrated with these original data (e.g., POS, lexical frequency).
Compiled data are extracted from the label files and stored in (compressed) tab-delimited plain text tables (ASCII). Entries are linked across tables with unique item (row) identifiers as proposed by [11]. Item identifiers contain pointers to recordings and label files.

4. Phonemic labeling and segmentation

By labeling and segmentation we mean 1. defining the phoneme (phoneme transcription) and 2. marking the start and end point of each phoneme (segmentation).

4.1. Procedure

The segmentation routine of an 'off-the-shelf' phone based HMM automatic speech recognizer (ASR) was used to time-align the speech files with a (canonical) phonemic transcription by using the Viterbi alignment algorithm. This produced an initial phone segmentation. The ASR was originally trained on 8 kHz telephone speech of phonetically rich sentences and deployed on downsampled speech files from the corpus. These automatically generated phoneme *labels* and *boundaries* were checked and adjusted by human transcribers (labelers) on the original speech files. To this end seven students were recruited, three males and four females. None of them were phonetically trained. This approach was considered justified since:
- phoneme transcriptions *without* diacritics were used, a derivation of the SAMPA set, so this task was relatively simple;
- naive persons were considered to be more susceptible to our instructions, so that more uniform and consistent labeling could be achieved; phonetically trained people are more inclined to stick to their own experiences and assumptions.

All labelers obtained a thorough training in phoneme labeling and the specific protocol that was used. The labeling was based on 1. auditory perception, 2. the waveform of the speech signal, and 3. the first spectral moment (the spectral center of gravity curve). The first spectral moment highlights important acoustic events and is easier to display and "interpret" by naive labelers than the more complex spectrograms. An on-line version of the labeling protocol could be consulted by the labelers at any time.
Sentences for which the automatic segmentation failed were generally skipped. Only in a minority of cases (5.5% of all files) the labeling was carried out from scratch, i.e. starting from only the phoneme transcription without any initial segmentation. The labelers worked for maximally 12 hours a week and no more than 4 hours a day. These restrictions were imposed to avoid RSI and errors due to tiredness.
Nearly all transcribers reached their optimum labeling speed after about 40 transcription hours. This top speed varied between 0.8 and 1.2 words per minute, depending on the transcriber and the complexity of the speech. Continuous speech appeared to be more difficult to label than isolated words, because it deviated more from the "canonical" automatic transcription due to substitutions and deletions, and, therefore, required more editing.

4.2. Testing the consistency of labeling

Utterances were initially labeled only once. In order to test the consistency and validity of the labeling, 64 files were selected for verification on segment boundaries and phonemic labels by four labelers each. These 64 files all had been labeled originally by one of these four labelers so within- as well as between-labeler consistency could be checked. Files were selected from the following speaking styles: fixed wordlist (W), fixed sentences (S), variable wordlist (W) and (variable) informal sentences (I). The number of words in each file was roughly the same. None of the chosen files had originally been checked at the start or end of a 4 hour working day to diminish habituation errors as well as errors due to tiredness. The boundaries were automatically compared by aligning segments pair-wise by DTW. Due to limitations of the DTW algorithm, the alignment could go wrong, resulting in segment shifts. Therefore, differences larger tan 100 ms were removed.

5. Results and discussion

The contents of the corpus at its first release are described in Tables 1 and 2. A grand total of 52 kWords (excluding filled pauses) were hand segmented from a total of 73 kWords that

were recorded (70%). The amount of speech recorded for each speaker varied due to variation in "long-windedness" and thus in the length of the informal stories told (which were the basis of the *Variable* text type). Coverage of the recordings is restricted by limitations of the automatic alignment and the predetermined corpus size.

In total, the ~50,000 words were labeled in ~1,000 hours, yielding an average of about 0.84 words per minute. In total, 200,000 segment boundaries were checked, which translates into 3.3 boundaries a minute. Only 7,000 segment boundaries (3.5%) could not be resolved and had to be removed by the labelers (i.e., marked as invalid).

The test of labeler consistency (section 4.2) showed a *Median Absolute Difference* between labelers of 6 ms, 75% was smaller than 15 ms, and 95% smaller than 46 ms. Pair-wise comparisons showed 3% substitutions and 5% insertions/deletions between labelers. For the intra-speaker re-labeling validation, the corresponding numbers are: a *Median Absolute Difference* of 4 ms, 75% was smaller than 10 ms, and 95% smaller than 31 ms. Re-labeling by the same labeler resulted in less than 2% substitutions and 3% insertions/deletions. These numbers are within acceptable boundaries [6] (sect. 5.2).

Regular checks of labeling performance showed that labelers had difficulties with:

1. The voiced-voiceless distinction in obstruents
2. The phoneme /S/ which was mostly kept as /s-j/; this was the canonical transcription given by CELEX
3. "Removing" boundaries between phonemes when they could not be resolved. Too much time was spent putting a boundary where this was impossible.

Using the compiled data tables fed into a PostgreSQL database allows to answer rather intricate questions. For instance, table 2 shows that, counter-intuitively, the articulation rates do not differ substantially between communicative speaking styles (I, R, T, S), but only for non-communicative styles (PS, W, Sy, Pr). Even fairly complicated questions, like comparing the durations of /m/ and /n/ in stressed syllables from spontaneous speech with respect to position in the word, ignoring sentence boundaries, becomes typing in a few commands, (e.g., /m/ vs. /n/ in ms, Initial: 71 vs. 63; Medial: 72 vs. 66; Final: 87 vs. 78).

6. Conclusions

A valuable hand-segmented speech database has been constructed in only 6 months of labeling, with 6 person-months of staff time for speech preparation and 1,000 hours of labeler time altogether. A powerful query language (SQL) allows comprehensive access to all relevant data.

This corpus is freely available and accessible on-line (http://www.fon.hum.uva.nl/IFAcorpus/). Use and distribution is allowed under the GNU General Public License (an Open Source License, see http://www.gnu.org). Direct access to an SQL server (PostgreSQL) is available as well as a simplified WWW front end. On-line, up-to-date, access to non-speech data is handled by a version management system (CVS).

7. Acknowledgments

Copyrights for this corpus, databases, and associated software belong to the Dutch Language Union (Nederlandse Taalunie). This work was made possible by grant nr. 355-75-001 of the Netherlands Organization for Scientific Research (NWO) and a grant from the Dutch "Stichting Spraaktechnologie". We thank Alice Dijkstra and Monique van Donzel of the NWO and Elisabeth D'Halleweijn of the Dutch Language Union for their practical advice and organizational support. Elisabeth D'Halleweijn also supplied the legal forms used for this corpus. Barbertje Streefkerk constructed the CELEX pronunciation list used for the automatic transcription.

8. References

[1] Bird, S., and Simons, G. "The open languages archives community", Elsnews 9.4, winter 2000-01, 3-5, 2001.

[2] Birney, E., Bateman, A., Clamp, M.E., and Hubbard, T.J. "Mining the draft human genome", Nature 409, 827-828, 2001.

[3] Cassidy, S. "Compiling multi-tiered speech databases into the relational model: Experiments with the EMU system", Proceedings of EUROSPEECH99, Budapest, 2239-2242, 2001.

[4] De Silva, V. "Spontaneous speech of typologically unrelated languages (Russian, Finnish and Dutch): Comparison of phonetic properties", INTAS proposal, 2000.

[5] Elenius, K. "Two Swedish speechdat databases - some experiences and results", Proceedings of EUROSPEECH99, Budapest, 2243-2246, 1999.

[6] Gibbon, D., Moore, R., and Winski, R. (eds.) "Handbook of standards and resources for spoken language systems", Mouton de Gruyter, Berlin, New York, 1997.

[7] Goedertier, W., Goddijn, S., and Martens, J.-P., "Orthographic transcription of the Spoken Dutch Corpus", Proceedings of LREC-2000, Athens, Vol. 2, 909-914, 2000.

[8] Grimm, J. and Grimm W. "Kinder- und Hausmaerchen der Brueder Grimm", 1857 (http://maerchen.com/)

[9] "Human Genomes, public and private", Editorial, Nature 409, 745, 2001.

[10] Matsui, T, Naito, M., Singer, H., Nakamura, A., and Sagisaka, Y, "Japanese spontaneous speech database with wide regional and age distribution", Proceedings of EUROSPEECH99, Budapest, 2251-2254, 1999.

[11] Mengel, A., and Heid, U., "Enhancing reusability of speech corpora by hyperlinked query output", Proceedings of EUROSPEECH99, Budapest, 2703-2706, 1999.

[12] Oostdijk, N., "The Spoken Dutch Corpus, overview and first evaluation", Proceedings of LREC-2000, Athens, Vol. 2, 887-894, 2000.

[13] Pols, L.C.W., "The 10-million-words Spoken Dutch Corpus and its possible use in experimental phonetics", Proceedings Int. Symp. on '100 Years of experimental phonetics in Russia', St. Petersburg, 141-145, 2001.

[14] "The principles of the International Phonetic Association", London, 1949.

[15] Williams, B., "A Welsh speech database: Preliminary results", Proceedings of EUROSPEECH99, Budapest, 2283-2286, 1999.

[16] Chan, D., Fourcin, A., Gibbon, D., et al. "EUROM – A spoken language resource for the EU", Proceedings EUROSPEECH95, 867-870, 1995.

[17] Burnage, G. "CELEX - A Guide for Users." Nijmegen: Centre for Lexical Information, University of Nijmegen. 1990.