Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 24 (2001), 15–26.

STRUCTURE AND ACCESS OF THE OPEN SOURCE IFA-CORPUS*

R.J.J.H. van Son & Louis C.W. Pols
{Rob.van.Son, Louis.Pols}@hum.uva.nl

Abstract

At the Institute of Phonetic Sciences (IFA) we have collected a corpus of spoken Dutch of 4 male and 4 female speakers, containing informal as well as read speech, plus lists of sentences, words, and syllables taken from the transcribed conversation text, and then spoken in isolation. This pertains to about 5.5 hours of speech. All this material is segmented and labeled at the phoneme level. This information plus all meta data are stored in a relational database which makes all material accessible through SQL. All information is freely available under the GNU General Public License. This material will also be used in INTAS project 915, in which a comparison will be made of phonetic properties in Dutch, Finnish and Russian. As an initial result we will present some durational and spectral data of full and reduced phoneme realizations.

Introduction

Speech and language research is quickly becoming a data-driven enterprise were large amounts of speech are needed to link the particulars of speech (e.g., coarticulation, reduction, prosody) to language (e.g., semantics, syntax) and vice versa. To service this need, more and more large speech databases are becoming available for speech research and commercial R&D (Gibbon et al., 1997, e.g., Cassidy, 2001; Elenius, 1999; Matsui et al., 1999; Oostdijk, 2000; Pols, 2001a; Williams, 1999). In our region we are fortunate to be involved in a process of collecting about 1,000 hours of spoken Dutch (Pols 2001a). This Dutch-Flemish project (Spoken Dutch Corpus, CGN; for more details see Oostdijk (2000) and http://lands.let.kun.nl/cgn/home.htm) will result in a highly accessible abundance of speech material transcribed at various levels, from many adult speakers, in various age groups, at three education levels, and in a variety of speaking styles. However, the collection of much speech material from single speakers under various conditions, is not foreseen in this project. Furthermore, none of the speech recordings will be phonemically segmented. In the presently popular variable-units concatenative synthesis it is customary to collect much speech material from a single speaker, but this is most of the time application-specific and in one (read) style only.

---

* This paper has been published in the proceedings of the IRCS workshop on Linguistic databases in Philadelphia, 11-13 December 2001. Parts of this paper have also been published in Van Son et al. (2001) and Pols and Van Son (in press).

Since we were interested in studying various reduction and coarticulation phenomena as a function of speaking style, word stress, sentence accent, position in the word, word frequency, and position of the word in the sentence (Pols 2001b), we decided to collect our own IFA-corpus. However, it would of course be foolish not to make good use of all experiences collected so far. So, we followed the CGN protocols as much as possible and used available software to ease orthographic transcription, to derive a phonemic transcription and a syllable split (CELEX), to perform forced phoneme alignment before doing manual adjustment, and to automatically extract part-of-speech tags and lemmas. All speech material is accessible via the user-friendly and powerful free speech signal processing package 'praat' developed at our institute (http://www.fon.hum.uva.nl/praat/) and is freely available upon request.

We also took great effort to put all non-speech data in an appropriate database structure, (http://www.fon.hum.uva.nl/IFAcorpus/) which makes it easily and freely accessible via a WWW interface.

Based on an inventarisation of our needs and the available funds, 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-off's involved in the construction of a corpus of hand-segmented speech to benefit future projects (e.g., the INTAS project (De Silva, 2000; Pols, 2001a).

Access and distribution of the available large databases are quickly becoming a problem. For instance, the complete Spoken Dutch Corpus (CGN Oostdijk, 2000; Pols, 2001a) will, for the time being, be distributed on about 175 CD-ROMs, making on-site management a real challenge. It is still not completely clear how best to access these data. Other corpora face the same problems (Cassidy, 1999; Chan et al., 1995; Elenius, 1999; Matsui et al., 1999; Williams, 1999).

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., Birney et al., 2001). Most queries are directed towards compiled data, not towards raw data. Many journals (e.g., Nature Editorial, 2001) 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 (Birney et al., 2001; Cassidy, 1999)

Table 1: Corpus contents (excluding empty 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	703	7307	10583	26021
G F/28	832	10944	799	10369	14664	35880
L F/40	640	8753	537	6954	10103	24792
E F/60	873	11246	711	8718	12931	31927
R M/15	655	7106	449	4581	6409	15642
K M/40	602	7667	400	4648	6612	15771
H M/56	675	8101	536	6446	9037	22559
O M/66	773	8237	287	2348	3398	8421
all	6128	73067	4492	51782	74702	187544

- Basic data should be available in compiled form
- Internet access is indispensable
- “Reviewed” user contributions should be stimulated and incorporated

1 Corpus content

Eighteen speakers (9 male and 9 female) participated in the recordings. Eight of them (4 male, 4 female) were selected for phonemic segmentation based on age and recording quality, and constitute the present IFA-corpus.

In Tables 1 and 2 the distribution of all segmented words per speaker and per speaking style are specified. All speech was recorded in a quiet, sound-treated room.

All audio-files were orthographically transcribed by hand according to the CGN protocol (Goedertier et al. 2000). A Dutch CELEX word list provided a pronunciation for most words as well as a syllable split-up, unknown words were hand-transcribed and added to the list.

Apart from the meta data, presently the following levels of transcription (plus segment boundaries) are available on separate tiers and can thus be the basis for subsequent analyses:

- the sentence level: reading text, orthographic transliteration
- the word level: orthography, realized and lexical phonemic transcription, POS, lemma, frequency
- the syllable level: realized and lexical, including lexical stress marks
- the demi-syllable level
- the phoneme level

Prominence marks as well as other prosodic transcriptions, via ToDI (http://lands.let.kun.nl/ todi) or otherwise, will be added later.

2 Corpus construction

2.1 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

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).

Speaker sex age	Style I	R	T	S	PS	W	Sy	Pr	All
N F 20	657	387	2418	2486	412	263	292	356	7271
G F 28	1859	1625	2732	2835	206	230	291	436	10214
L F 40	887	466	2117	2072	423	239	274	345	6823
E F 60	929	1172	2534	2744	214	262	315	407	8577
R M 15	118	321	1321	1430	439	233	268	423	4553
K M 40	534	433	1340	1332	0	249	275	415	4578
H M 56	266	656	1991	2071	435	261	286	450	6416
O M 66	0	1169	0	0	425	193	120	437	2344
All	5250	6229	14453	14970	2554	1930	2121	3269	50776
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	

speakers (9 male, 9 female) completed both recording sessions. All speakers were mother-tongue speakers of Dutch 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).

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.

2.2 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*" (IPA, 1949), and on a translation of the fairy tale "*Jorinde und Joringel*" (Grimm and Grimm, 1857). 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 (Burnage, 1990) 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.

2.3 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*).

2.4 Speech preparation, file formats, and compatibility

The corpus discussed in this paper is constructed according to the recommendations of (Gibbon et al., 1997; Goedertier et al., 2000). Future releases will conform to the Open Languages Archives (Bird and Simons, 2001). 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 (Goedertier et al., 2000). 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.fon.hum.uva.nl/praat).

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, lemma, 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 (Mengel and Heid, 1999). Item identifiers contain pointers to recordings and label files.

3 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).

3.1 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. Labeling speed from scratch was about half the speed for pre-aligned speech. 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.

3.2 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 than 100 ms were removed.

4 Results

The contents of the corpus at its first release are described in Tables 1 and 2. A grand total of 50 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 was 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 monetary cost of the automatic and manual alignment combined (excluding VAT) was DFl 74,000 in total (33,597 Euro). This translates to around DFl 1.40 per word (0.65 Euro/word) and DFl 0.37 per boundary (0.17 Euro/boundary). The total staff time needed to prepare and transliterate the speech and manage the automatic

Table 3. Occurrence of surface plural /-n/ in nouns and verbs for different styles. Percentages are not affected by excluding cases where the next word starts with a vowel. The differences are significant ($X^2 = 307$, DoF = 4, $p < 10^{-5}$)

Style	/@n/	/@/	All	% /@n/
I	1	304	305	0.3
R	13	236	249	5.2
T	180	372	552	33
S	203	340	543	37
PS	62	19	81	77
All	459	1271	1730	36

pre-alignment and human labelers was around 6 person-months (half of which was not included in the budget quoted above).

The test of labeler consistency (section 3.2) 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 (Gibbon et al., 1997; sect. 5.2).

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

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

5 Access and SQL querying

Speech and language corpora are huge stores of data. The question is how these massive bodies of data can become useful for research. Essentially, there are two major approaches. First, people will try to determine what is in the store, i.e., exploring and counting whatever phenomenon they are interested in. That is, they want descriptive statistics on subsets of the corpus. A lot of very advanced research can be done on compiled statistics of corpus data (Birney et al., 2001). However, there will always be users that need access to the raw data itself: recordings, analysis and annotation files. These users need powerful methods for selecting the relevant subsets of the corpus. Both these approaches to corpus use are implemented in the IFA corpus. Fundamental to both approaches is the ability to intelligently query the stored information.

Therefore, to make a corpus usable, it must be possible to query it efficiently. For many purposes and database types there exist specialized languages which allow to extract the relevant information (e.g., Cassidy, 1999). The most general used and best understood database type is the relational database and it's basic query language is SQL. There exist extremely efficient and reliable off-the-shelf open source implementations of relational databases and SQL, that can also be used over the internet. As many (if not most) query languages can be mapped onto SQL (it is *complete* as a query language, e.g., Cassidy, 1999), we decided to store all our data in a relational database (i.e., PostgreSQL) and use SQL as the query language. This solves many problems of storage, access, and distribution.

Although access to our corpus and database by way of SQL queries is possible over the internet, this cannot be granted directly to anonymous users because of

Table 4. Occurrence of surface plural /n/ in nouns and verbs for words with low (•0.0001) and high (>0.0001) frequency of occurrence in read speech (T,S). Percentages are not affected by excluding cases where the next word starts with a vowel. The differences are significant (X^2 = 14,42, DoF = 1, p < 0.0002) (note: 5 words had no frequency data and were omitted)

Freq.	/@n/	/@/	All	%/@n/
Low	176	244	420	42
High	204	466	670	30
All	380	710	1090	35

Table 5. Schwa epenthesis between /l/ or /r/ and a following syllable-final /kmpfvbxX/. The differences are not significant (X^2 = 3.62, DoF = 7, p > 0.05)

Style	Epenthesis	None	All	% /@/
I	10	49	59	17
R	14	49	63	22
T	20	117	137	15
S	24	121	145	17
PS	6	13	19	32
W	6	24	30	20
Sy	5	22	27	19
Pr	14	48	62	23
All	99	443	542	18

security concerns. Therefore, we added a WWW front-end to the corpus and database. This allowed us to simplify access by automatically generating complex SQL queries and direct links to the relevant files. Annotations, transcriptions and other human derived data are stored in a version system (CVS) that allows collaborative updates and version histories over the internet. This system was indispensable during corpus generation as this was done at separate locations.

5.1 Query examples

With the implemented data structure and a powerful query language SQL, it is possible to answer rather intricate questions such as in the following examples (taken from our Web Interface manual page on: http://www.fon.hum.uva.nl/IFAcorpus)

- what is the average articulation rate per sentence, expressed in number of syllables or phonemes per second, for these various speaking styles? See Table 2.
- to what extend is the surfacing of the plural /n/ in nouns and verbs a “reading” artifact. See Table 3.
- the same question, but now for low and high frequency words in read texts and sentences. See Table 4
- is the occurrence of schwa epenthesis between /l/ or /r/ and syllable final (non-alveolar/palatal) obstruents sensitive to the style of the speech. See Table 5.
- what is the corrected means duration of all intervocalic consonants in polysyllabic, non-high-frequent words, not at sentence boundaries, as a function of the within word position and the syllable stress, both in read as well as in spontaneous speech? See Table 6.
- what are the average vowel positions in the F1/F2 space in different speaking style conditions? See Figure 1.

Table 6. Corrected means duration in ms of intervocalic consonants (nasals, fricatives, stops, and glides),word freq. < 0.001, as a function of position in the word, syllable stress (+/-), and spontaneous or read speech. Italic numbers: phoneme counts.

	Spontaneous		Read		Total
Stress	+	–	+	–	count
Initial	71 *202*	59 *96*	73 *715*	68 *285*	*1298*
Medial	63 *295*	61 *810*	69 *837*	63 *2586*	*4528*
Final	86 *20*	74 *94*	74 *75*	67 *317*	*506*

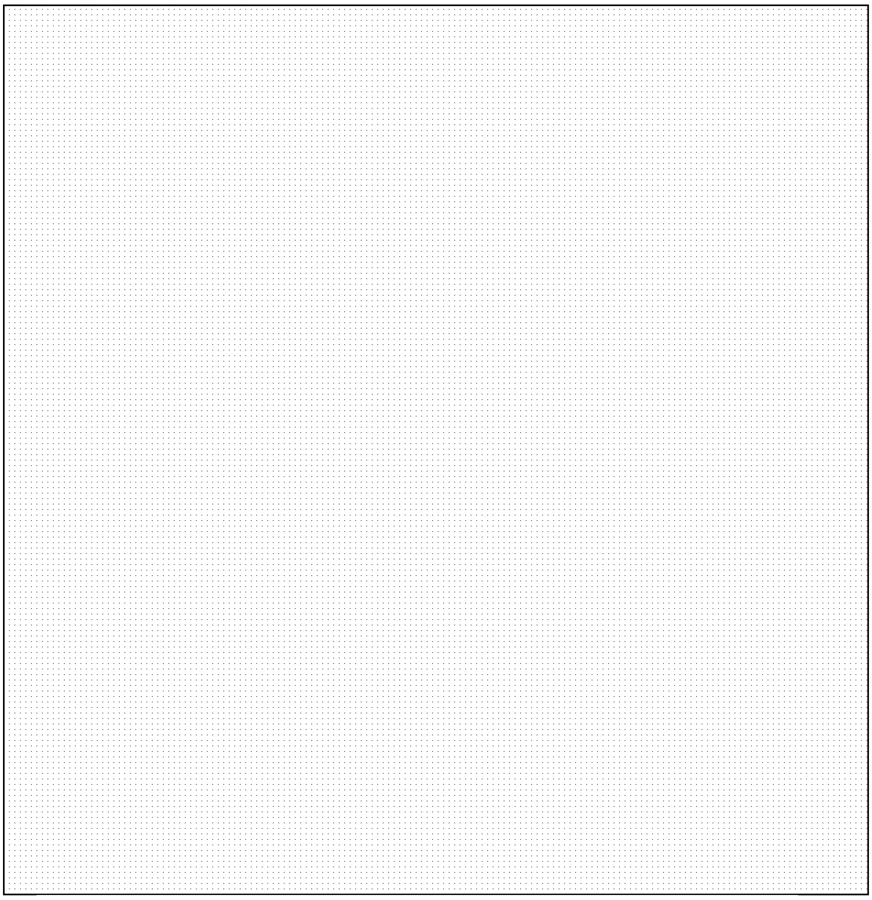

Fig. 1 Average vowel formant positions for one female speaker in three speaking style conditions. For more details, see text.

5.2 Discussion of query examples

Somewhat to our surprise the articulation rates do not differ much between the first four communicative speaking styles, of which the first two represent conversational speech and the next two read speech (Table 2, last two rows). The final four non-communicative speaking styles indeed do show substantially lower rates.

The plural /-n/ ending in Dutch Nouns and Verbs is always written in the orthography, but generally not spoken in informal speech. Therefore, we can infer that this might be an example of orthographic interference. It is clear from Table 3 that pronouncing the plural /n/ is indeed largely confined to read speech.

However, this example also demonstrates the danger of blind trust in global corpus statistics. To ease the task of the labelers, we have transcribed all read speech with final /@n/ where the canonical CELEX word-list used final /@/. On the other hand, informal and retold speech (*I* and *R*) were transcribed without the plural /-n/ ending. So there is a strong initial transcription bias against final /@n/ in informal and retold speech and for surfacing plural /-n/ in read speech.

If the surfacing of plural /-n/ is indeed a reading (i.e., task) artifact, then we can expect a sensitivity to word frequency. That is, common words should have less plural /-n/ endings than rare words. This is indeed found when the query is repeated for read speech and high- and low-frequency words (Table 4, i.e., with fixed transcription bias). Therefore, we can indeed state that surfacing of plural /-n/ endings is largely a reading artifact, i.e., orthographic interference.

In contrast, no systematic effects of speaking style could be found for schwa epenthesis after /l/ or /r/ (Table 5).

For the data in Table 6 a more complex analysis was required, we implemented a so-called corrected means analysis (van Santen 1992) which takes into account the unequal distribution of values in each cell. It is worth noting the long duration for the consonants in stressed syllables in word final position. Unfortunately the number of observations is rather low for these cells.

The durational measurements, as presented in the above tables, could be derived directly from the segment boundaries. But of course other parameters can rather easily be derived within 'praat', such as pitch, formant frequencies, intensity, or center of gravity. In Fig. 1 below we present the average vowel formant positions in F_1-F_2 for three speaking style conditions, namely:

- at least 4 repetitions of clearly pronounced vowels in isolation or in spelled letters of the alphabet. See filled circles in Fig. 1;
- vowels taken from read sentences. See open triangles in Fig. 1;
- vowels taken from an informal story told by this female speaker face-to-face to an interviewer. See open circles in Fig. 1.

These data are from one of the four female speakers in this IFA-corpus. All vowel segments per condition are used for this analysis, but for the last two conditions only realizations from multi-syllabic words and in lexically stressed position were used. The schwa was always excluded. The segment selection as well as the formant measurements (at the midpoint in each vowel segment) were done fully automatically. For large amounts of data this is the only possible way. However, unavoidably this might introduce some inconsistencies and errors. For instance, the average data in Fig. 1 are sometimes based on only 3 realizations (for the rare vowel /ø/, presented in the figure with the SAMPA symbol '2'), sometimes on as many as 127 (for the vowel /e/ in read sentences). Furthermore, not all formant measurements may be fully reliable. For instance, the standard deviation for the first formant measurements of the vowels /A/ and /a/ in the informal speaking style is rather high, just as some of the second formant measurements for some other vowels, which may have to do with effects of reduction, coarticulation, diphtongization, or perhaps even labeling errors. But despite these imperfections, this figure nicely illustrates for 'real speech data' the large spread of the vowel space if the utterances are clearly spoken, as well as the substantially reduced, but still easily recognizable, vowel triangle for more conversational speech. Actually we performed similar measurements for the unstressed realizations as well (not shown here), and found of course much more centralization in those conditions.

Conclusion

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). In the near future we will extend our analyses of this highly interesting speech material and we will compare the data for Dutch with those for Finnish and Russian. We will also add prosodic annotations to make this material even more useful.

Acknowledgements

Diana Binnenpoorte and Henk van den Heuvel of SPEX designed, organized, and managed the segmentation and labeling, including the selection and training of the labelers, and supplied the section on the labeling speed and consistency. They also co-authored an earlier paper of which parts are reproduced here (Van Son et al., 2001). Copyrights for the IFA corpus, databases, and associated software are with the Dutch Language Union (Nederlandse Taalunie). This work was made possible by grant nr 355-75-001 of the Netherlands Organization of Research (NWO) and a grant from the Dutch "Stichting Spraaktechnologie". We thank Alice Dijkstra and Monique van Donzel of the Netherlands Organization for Scientific Research (NWO) and Elisabeth D'Halleweijn of the Dutch Language Union (Nederlandse Taalunie) for their practical advice and organizational support. Elisabeth D'Halleweijn also supplied the legal forms used for this corpus. Barbertje Streefkerk constructed the CELEX word list used for the automatic transcription.

References

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

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

Burnage, G. (1990). "CELEX - A Guide for Users." Nijmegen: Centre for Lexical Information, University of Nijmegen.

Cassidy, S. (1999). "Compiling multi-tiered speech databases into the relational model: Experiments with the EMU system", Proceedings of EUROSPEECH99, Budapest, 2239-2242.

Chan, D., Fourcin, A., Gibbon, D. (1995). "EUROM - A spoken language resource for the EU", Proceedings EUROSPEECH'95, 867-870.

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

Editorial (2001). "Human Genomes, public and private", Nature 409, 745.

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

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

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

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

IPA (1949). "The principles of the International Phonetic Association", London.

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

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

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

Pols, L.C.W. (2001a). "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.

Pols, L.C.W. (2001b). "Acquiring and implementing phonetic knowledge", Proc. Eurospeech 2001, Aalborg, Denmark, Vol. 1, K-3-K-6.

Pols, L.C.W., and Van Son, R.J.J.H. (in press). "Accessing the IFA-corpus".

Van Santen, J.P.H. (1992). "Contextual effects on vowel duration", Speech Communication 11, 513-546.

Van Son, R.J.J.H., Binnenpoorte, D., van den Heuvel, H. and Pols, L.C.W. (2001). “The IFA corpus: a phonemically segmented Dutch Open Source speech database”, Proc. EUROSPEECH 2001, Aalborg, Denmark, Vol. 3, 2051-2054.

Williams, B., (1999).”A Welsh speech database: Preliminary results”, Proc. of EUROSPEECH99, Budapest, 2283-2286.