DUTCH CORPUS
OF
PATHOLOGICAL AND NORMAL
SPEECH (COPAS)

Gwen Van Nuffelen
Marc De Bodt
Antwerp University Hospital
Department ORL, Head and Neck Surgery
Belgium

Catherine Middag
Jean-Pierre Martens
Ghent University
Department Electronics & Information Systems
Belgium

Content

Introduction

The Dutch Corpus of Pathological and Normal Speech (COPAS) was constructed within the framework of the project Speech Algorithms for Clinical and Educational applications (SPACE) which was granted by the Flemish Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) (contract SBO/40102). This 4-year project started in March 2005. One of the goals of SPACE was to develop a reliable, speech technology based assessment tool for pathological speech, more precisely an objective intelligibility assessment (Van Nuffelen et al 2009, Middag et al 2008). The development of this objective intelligibility assessment is based on recordings of the standardized and reliable perceptual Dutch Intelligibility Assessment (DIA). Therefore the majority of the samples in this corpus are recordings of the DIA but the corpus also contains a variety of other samples like reading passages, isolated sentences and recordings of spontaneous speech. The corpus contains both samples of speakers with and without a speech disorder. The included speech disorders are: dysarthria, voice disorders, cleft, functional or organic articulation disorders, laryngectomy, glossectomy and impaired speech secondary to hearing impairment.

The corpus can be used to train speech language pathology students, allowing them to get familiarized with the several perceptual features of pathological speech and to develop or ameliorate speech technology based assessment and treatment tools for pathological speech. The authors intend to gradually enlarge the corpus during the following years.

1. General information

General information about the corpus can be found on in the folder COPAS/INFO. The information is split into two parts.

Subfolder MATERIAL contains all information about the text materials that were used during the recordings.

Subfolder PATIENTS contains all information about the speakers that have been recorded.

2. Speaker information

The speakers recorded in COPAS belong to 8 distinct pathological categories, which are shown together with their abbreviation and cardinality (number of speakers of that category) in Table 1.

Table 1

Speakers	**N**
Normal (N)	122
Dysarthria (D)	75
Hearing impairment (H)	29
Laryngectomy (L)	30
Cleft (C)	38
Articulation disorders (A)	17
Voice disorder (V)	7
Glossectomy (G)	1
Total	319

The speaker information in COPAS/INFO/PATIENTS is organized per pathology. Per category there is an excel file with one or two sheets. The first sheet is called ‘recordings’ and contains the following information of each speaker:

- ID: A speaker identification code consisting of a single character (referring to the pathology) and a speaker identification number
- First: If the speaker was recorded more than once, this column contains the ID that was used for the first recording of this speaker
- Ttf: If this is not the first recording of the speaker, this column contains the time (in months) that passed between the first and the actual recording.
- Gender: The speaker gender (M/F)
- Age: The age of the speaker (in years) at the time of the recording. If the age is unknown, it is marked as “unk”.
- MIC: The microphone that was used for the recording (sony, shure). If one doesn’t know, the field is marked with “unk” (unknown)

- A, B, C,..,SV: The different subtests (see below) that were recorded (see below for more details). If the field is empty, the subtest was not performed, if it is not empty the field provides information about the text material that was read by the speaker.
- I_A, I_B, I_C, I: Perceptual intelligibility scores available for the speaker on the basis of the recording. These scores will be explained further on.

The second sheet, if present, is called ‘pathologies’. It repeats the first four fields of the ‘recordings’ sheet, and for the rest it contains details about the type of pathology and/or the severity of that pathology. This information is differs from one pathological class to the other.

The excel file **overview.xls** provides a summary of the number of recorder speakers per speech pathology and the number of speakers per category that participated in the different tests. It contains two tables: one with the total number of **speaker IDs** that were used and one with the number of **different speakers** that were actually being recorded (the doubles are not counted here).

3. Test information

Several tests were recorded, but not every speaker participated in every test. Table 1 presents an overview of all tests (represented by their acronym) with the number of performers per pathology (some performers repeated the same test more than once, but they are counted only once here). The table also indicates for which tests an annotation (see below) is available.

Table 2

	N	**D**	**H**	**L**	**C**	**A**	**V**	**G**	**total**	**annotated**
total	122	75	29	30	38	17	7	1	319	
DIA	122	75	29	30	38	3	7	1	305	Yes
ART	0	0	0	0	0	14	0	0	14	Yes
SV	84	56	24	7	0	0	4	1	176	No
DKR	83	55	24	7	0	0	3	1	173	No
FT	81	52	24	7	0	0	3	1	168	No
S1	82	50	24	7	0	0	3	1	167	Yes
S2	82	50	24	7	0	0	3	1	167	Yes
TM	82	49	26	15	0	0	6	1	179	Yes
T	81	25	9	0	0	0	0	0	115	Yes
SS	0	20	1	0	0	0	4	0	25	No
SSS	80	21	10	17	0	0	0	0	128	No

The subsequent paragraphs review the different tests and the text materials that were used in the context of these tests.

Dutch intelligibility assessment (DIA)

The DIA (De Bodt et al 2006) was constructed to measure intelligibility of the speaker at the phoneme level and to give the clinician an overview of the kind of segmental articulatory errors the speaker makes. Every speaker that can read (few exceptions) participated in this test and read a series of 50 consonant-vowel-consonant words which are divided in 3 subtests: subtest A (19 words) assesses all the initial consonants, subtest B (15 words) all the final consonants and subtest C (16 words) all the medial vowels and diphthongs of Dutch. There are 35 versions of each subtest. This gives 3 x 35 text files in folder COPAS/MATERIAL/DIA. The fields A, B and C in the 'recordings' sheet of the speaker information file contain the versions that were read by a given speaker.
The recorded material is judged by a human listener and the perceptual scores representing phoneme intelligibilities (De Bodt et al. 2006) are included in the speaker information file as columns I_A, I_B, I_C and I. The first three represent the number of correctly perceived target phonemes in the three subtests, and I is the total phoneme intelligibility (in %) that is derived from these three subscores.

Text (T)

In this test, the speaker reads one of a set of eleven different text passages which all have a reading difficulty level AVI 7 or 8. The passages are available in folder INFO/MATERIAL/T, and the field T in the speaker information contains the number of the passage (1..11) that was read by the speaker.

Text Marloes (TM)

This text is a standardized text with a balanced representation of Dutch phonemes. It is often used in clinical practice. The text is available in TM.txt in folder COPAS/MATERIAL/TM, and the field TM in the speaker information file is set to 1 if the test was recorded.

Articulation assessment (ART)

Children with insufficient reading skills were assessed by means of the Logo-Art articulation assessment (Baarda et al 2001) which is a picture naming test. There are seven subtests each targeting a different phoneme position and category. The subtests refer to: initial consonant (CI), final consonant (CF), medial consonant (CM), vowels (V), diphthongs (D), initial consonant clusters (CCI) and final consonant clusters (CCF). The target words used in the different subtests are provided as text files (e.g. CCF.txt) in folder COPAS/ INFO/MATERIAL/ART. The fields CI, .., CCF in the speaker information file are set to 1 if the corresponding subtests were recorded. The speakers participating in the ART not always performed all seven subtests.

Since the children had to name pictures, they did often gave alternative names for the targeted name (e.g. "twee" (two) instead of "tweeling" (twins)). As marking these alternatives in the text files on COPAS/ INFO/MATERIAL/ART would mean that these files continuously change as more data are recorded, we have opted for a solution in which the target word lists remain fixed but the alternatives appear in the annotation files of the recorded sessions (see section 5).

Diadochokinetic rate (DKR)

This is one of the tasks of the Motor Speech Profile (Kay Elemetrics). The speaker is instructed to repeat /p/-/p/-/p/ for at least 8 seconds. If the test was recorded, the DKR field in the speaker information file is set to 1.

Formant transition (FT)

This is also one of the tasks of the Motor Speech Profile (Kay Elemetrics). The speaker is asked to alternate /i/ (the vowel of 'tien') and /u/ (the vowel of 'doel') for at least 6 seconds. If the test was recorded, the field FT in the speaker information file is set to 1.

Sentence 1 (S1)

The speaker is asked to imitate the sentence 'Wil je liever de thee of de borrel?' with the appropriate intonation and stress according to the example in the Motor Speech Profile. The sentence is available in S1.txt in folder COPAS/MATERIAL/S1. If the test was recorded, the field S1 in the speaker information file is set to 1.

Sentence 2 (S2)

The speaker is asked to read or repeat the sentence 'Na nieuwjaar was hij weeral hier'. The sentence is available in S2.txt in folder COPAS/MATERIAL/S2. If the test was recorded, the field S2 in the speaker information file is set to 1.

Spontaneous speech (SS)

For these samples the speakers were asked to tell something about their work, hobby, family and so forth. If the test was recorded, the field SS in the speaker information file is set to 1.

Semi spontaneous speech (SSS)

The speakers were instructed to tell a story based on a randomly selected sequence of pictures (Color Cards). If the test was recorded, the field SSS in the speaker information file is set to 1.

Sustained vowel (SV)

The speaker is instructed to sustain the vowel /a/ (like in 'maan') for at least 5 seconds at habitual loudness and pitch. If the test was recorded, the field SV in the speaker information file is set to 1.

4. Audio samples

All recordings were made in a quiet clinical setting without sound treated box. Two microphones were used: a Sony ECM-717 lying on the table, with mouth-microphone distance of about 30 cm, and a Shure headset WH20-QTR. The reason for using two different types of microphones is historical.

The speech was recorded by means of a mini-disc (Sony MZR700) and later transferred to a notebook. The transfer was done with a freely available wave editor (Audacity®). It is performed in such a way that one audio sample is created for each subtest being recorded. Each audio sample is stored in Windows wav format: the sampling rate is 16000 Hz and the encoding is 16 bit linear PCM.

The audio samples are organized per test. All the audio samples recorded for a particular test are in one subfolder, called COPAS/SAMPLES/xxx, with xxx representing the acronym of the test (as listed in Table 2).

5. Annotations

The annotations were performed with the open source program PRAAT (Boersma et al. 2006) and are stored in TextGrid files. Each file is organized in tiers, with each tier comprising time intervals with a piece of text attached to them. The TextGrid files can be visualized together with the speech in PRAAT.

The first tier of the TextGrid file always represents the target text that was read, whereas the second tier represents what the annotator perceived. This perception is annotated as an *orthographic transcription* or a *transliteration* of the speech.

In the case of ART, the speaker may have used a valid alternative for the target word (see section 3). In that case, the first tier contains this alternative, but to indicate that this is an alternative, it is put between brackets.

In the case of DIA, the transliteration only differs from the target text in the phoneme that was actually tested in the recorded subtest (e.g. the initial consonant), even if the rest of the word is not perceived as implied by the target word either.

Time intervals containing silence, noise or interruptions are marked with ‘nsp’, which stands for ‘no speech’.

The annotation of a file is in the same folder as the audio sample on which it is based.

6. Filenames

The name of each audio file (extension **.wav**) and annotation file (extension **.TextGrid**) of the same speaker will start with a single character referring to the pathology, a speaker identification code (pathology category + speaker ID), an underscore, a string referring to the subtest being recorded, and optionally (for the DIA and the T test only) the version of the subtest that was used. The subtest strings are presented in Table 3. From that table it follows that the filename A12_A17.wav thus contains version 17 of subtest A of the DIA test, as it was recorded from speaker 12 listed in the Articulation Disorder speaker information file.

Table 3

A	Subtest A of the DIA	
B	Subtest B of the DIA	
C	Subtest C of the DIA	
T	Text	
TM	Text Marloes	
CI	Initial consonants	(part of the articulation assessment)
CF	Final consonants	(part of the articulation assessment)
CM	Medial consonants	(part of the articulation assessment)
V	Vowels	(part of the articulation assessment)
D	Diphthongs	(part of the articulation assessment)
CCI	Consonant clusters initial	(part of the articulation assessment)
CCF	Consonant clusters final	(part of the articulation assessment)
DKR	Diadochokinetic rate	
FT	Formant transition	
S1	Sentence 1	
S2	Sentence 2	
SS	Spontaneous speech	
SSS	Semi spontaneous speech	
SV	Sustained vowel	

7. References

Boersma P, Weenink D (2006). Praat: doing phonetics by computer (Version 4.4.34) [Computer program]. Retrieved October 19, 2006, from http://www.praat.org/

Baarda D. de Boer-Jongsma N., Haasjes-Jongsma W. (2001). Logo-Art: Articulatieonderzoek. Baert: Ternat.

De Bodt M, Guns C , Van Nuffelen G (2006). NSVO: handleiding. Vlaamse Vereniging voor Logopedie: Herentals.

Middag C, Van Nuffelen G, Martens JP, De Bodt M (2008). Objective intelligibility assessment of pathological speakers. Proceedings Interspeech 2008, Brisbane.

Van Nuffelen G, Middag C, De Bodt M, Martens JP (2009). Speech technology based assessment of phoneme intelligibility in dysarthria. International Journal of Language and Communication Disorders 44, 716-730.