

# Sense Meets Nonsense

a dual-layer Danish speech corpus for perception studies

**Christiansen, Thomas Ulrich; Henrichsen, Peter Juel**

*Published in:*
8th International Conference on Language Resources and Evaluation

*Publication date:*
2012

# Sense Meets Nonsense – a dual-layer Danish speech corpus for perception studies

**Thomas Ulrich Christiansen[1] and Peter Juel Henrichsen[2]**

[1]Centre for Applied Hearing Research,
Technical University of Denmark,
Ørsteds Plads 1, Building 352,
DK-2800 Lyngby, Denmark
[2]Center for Computational Modelling of Language,
Copenhagen Business School
Dalgas Have 15,
DK-2000 Frederiksberg, Denmark
[1]tuc@elektro.dtu.dk, [2]pjh.isv@cbs.dk

**Abstract**

In this paper, we present the newly established Danish speech corpus PiTu. The corpus consists of recordings of 28 native Danish talkers (14 female and 14 male) each reproducing (i) a series of nonsense syllables, and (ii) a set of authentic natural language sentences. The speech corpus is tailored for investigating the relationship between early stages of the speech perceptual process and later stages. We present our considerations involved in preparing the experimental set-up, producing the anechoic recordings, compiling the data, and exploring the materials in linguistic research. We report on a small pilot experiment demonstrating how PiTu and similar speech corpora can be used in studies of prosody as a function of semantic content. The experiment addresses the issue of whether the governing principles of Danish prosody assignment is mainly talker-specific or mainly content-typical (under the specific experimental conditions). The corpus is available at http://amtoolbox.sourceforge.net/pitu/.

**Keywords:** speech corpus, Danish language, nonsense syllables, prosodic structure, corpus-based spoken language analysis

## 1. Introduction

In many current models of human language processing, the speech decoding process is described as a series of analytical stages, beginning at the psychoacoustic perception level and gradually abstracting away from the physical manifestation of the speech signal through stages of phonetic, phonological, morphological, syntactic, semantic, and finally pragmatic processing. Little is known about the extent to which such stratified description models are indeed paralleled by discrete sub-processes in the brain. One way to approach questions of the "modularity of mind" is to prepare a combined source of two (or more) dichotomic data sets. The data sets should keep constant as many mutual conditions as possible while varying one parameter only, viz. the point of attachment to the stratified model. In the corpus presented here under the name PiTu, we wanted to keep constant the perception situation (including location, time, recording equipment, individual talker, etc) while systematically varying the amount of semantic-pragmatic interpretation involved in solving the reproduction task. Our talkers were thus instructed to first repeat six series of nonsense syllables (65 in all), and immediately after to read aloud a list of sentences (selected from 10 authentic text sources). More details are given below. See also (Christiansen and Henrichsen, 2011).

### 1.1. Language is meaningful - so why use nonsense syllables?

Early in the 20th century Harvey Fletcher (e.g. (Fletcher, 1920)) investigated speech intelligibility of nonsense-syllables in order to maximise perceptual throughput of telephone lines. The idea behind using nonsense-syllables was that context effects and idiosyncratic effects from "meaningful speech" would not have to be controlled for in the experimental set-up. The perhaps most remarkable results from this research was the Articulation Index (AI), which predict speech intelligibility based on frequency specific signal-to-noise-ratios and importance weights (e.g. (Fletcher and Galt, 1950)). Even today telephone lines carry the frequency most important for intelligibility as predicted by AI. (Miller and Nicely, 1955) complemented Fletcher's experiments by high-pass and low-pass filtering nonsense-syllable and examine intelligibility in the presence of background noise. They analysed the results by means of confusion matrices and found that the distinctive phonetic features voicing, manner and place of articulation are perceived rather differently. Voicing can be recognised even with only narrow frequency bands available whereas place of articulation requires broader bands. Recently, (Christiansen and Greenberg, 2012) elaborated this finding and showed that while the spectral integration function for consonants in nonsense-syllables is linear, the underlying functions for voicing, manner and place of articulation

are compressive, linear and expansive, respectively. Further, they suggested that the distinctive phonetic feature are processed hierarchically. Clearly speech perceptual research involving nonsense-syllables, i.e., without meaning only addresses processing in the early stages of perception. In order to understand the later processing stages it is necessary to study speech with meaning. (Bronkhorst et al., 1993) suggested a model accounting for co-articulatory effects in the perception of consonant-vowel-consonant (CVC) syllables. Later, using other speech material, (Bronkhorst et al., 2002) suggested a model to account for perceptual context effects in sentences, i.e., meaningful speech. The speech material for the sentences was not spoken by the same talkers as the speech material for the individual phonetic segments in the CVC study. The material presented here contains both sentences and nonsense-syllables spoken by the same talkers.

## 2. Speech material

The PiTu corpus consists of two parts nonsense syllables and sentences with meaning. The nonsense syllables were prompted – the sentences were read from a sheet of paper. In the following sections we describe the details of the materials and their recording.

### 2.1. The nonsense syllables

The Danish consonants recorded in the present study correpond to the phonemes /ptkbdgfsvmnrlhʃjw/[1] roughly corresponding to the following phones in IPA-notation (IPA, 1999) [pʰtˢkʰbd̥g̊fsʊmnʁlhʃjw]. Note that the two approximants /j/ and /w/ were included in the recording as if they were consonants[2] .

Consonants were followed by one of three long vowels /iau/ corresponding to vowel qualities designated by IPA-symbols [iæu]. This first consonant-vowel (CV) syllable was stressed. Some combinations of consonants and vowels coincide with Danish words. In order to dissociate meaning from all syllables a second unstressed /tu/-syllable was added. So the recorded nonsense syllables consisted of four speech sounds a consonant and a vowel followed by /tu/. We refer to these syllables as CV-triplets.

To keep talkers alert six fillers with unstressed second syllable /ta/ ([tæ] in IPA notation) was incorporated into the material (/ʃata/ /lata/ /wita/ /mita/ /ruta/ /juta/). Eight additional /v/-syllables was included, since we speculate that /v/ is articulated with a higher degree of variability than the other consonants and plan to investigate this speculation elsewhere.

Five lists with each eleven syllable pairs and one list with ten syllable pairs were constructed, i.e., a total of 65 syllable pairs. These list were made up from three types of syllable pairs: 1) all combinations of seventeen consonants and three vowels (= 51 CVtu syllable pairs) 2) six fillers (= 6 CVta syllable pairs) and 3) eight additional CVtu syllables with consonant /v/ (=8 CVtu syllable pairs). These syllable pairs were transcribed and randomly distributed across the six lists as shown in Table 1.

| List 1 | List 2 | List 3 | List 4 | List 5 | List 6 |
|---|---|---|---|---|---|
| pa:tu | pi:tu | pu:tu | ka:tu | ki:tu | ku:tu |
| ru:tu | nu:tu | mi:tu | ma:tu | na:tu | ni:tu |
| vi:tu | va:tu | li:tu | vu:tu | mu:tu | la:tu |
| ʃa:ta | ju:ta | ru:ta | wi:ta | la:ta | mi:ta |
| ti:tu | ta:tu | da:tu | bu:tu | tu:tu | ba:tu |
| vu:tu | fi:tu | fu:tu | fa:tu | si:tu | sa:tu |
| ha:tu | ra:tu | ri:tu | lu:tu | hu:tu | hi:tu |
| vi:tu | vi:tu | vu:tu | vu:tu | vi:tu | vu:tu |
| wa:tu | ʃu:tu | wi:tu | ʃi:tu | ja:tu | ga:tu |
| su:tu | ji:tu | ʃa:tu | vi:tu | wu:tu | ju:tu |
| bi:tu | du:tu | gi:tu | gu:tu | di:tu | |

Table 1: The six list of nonsense-syllables recorded in the PiTu corpus

#### 2.1.1. Recording procedure

The recordings were carried out in two stages. The aim of the first stage was to produce a CD, which could be used in the second stage. This CD contains sound recordings of nonsense syllables as shown in Table 1.

The second stage consisted in recording talkers repeat the content of the CD from the first stage. The recordings from the second stage is the topic of the present paper while the recordings from the first stage is merely used as prompting material.

In the first stage the authors were recorded speaking each item from Table 1 three times in succession with the neutral sentence intonation contour for Danish (falling). At the beginning of each recording the authors uttered the Dansih phrase “Nu bliver der sagt” (English: “Now this will be said”). The best of the two recordings was used to produce the CD.

The nonsense syllables were put on the CD with six tracks, each of which corresponds to a column in Table 2.1. such that each track starts with the prompting sentence “Nu bliver der sagt” immediately followed by the first nonsense syllable repeated three times. We refer to these three utterances of the nonsense syllables as a triplet. Subsequent triplets were preceded by 4 seconds of silence. This allows for the talkers to repeat the triplet from the CD.

In the second stage recordings were carried out in the small anechoic chamber at the Technical University of Denmark (Ingerslev et al., 1968) using a low noise 1-inch B&K 4179 microphone with a B&K 2660 preamplifier attached to a SoundDevices 722 harddisk recorder. The microphone power supply was a B&K 2807. The microphone was mounted on a stand no less than 1 meter from the mouth of the talker. The talker was seated in a desk chair facing the microphone. The system was calibrated with a B&K 4239 calibrator so

[1] We adopt the common practice of denoting phonemes between // and phones in [ ]

[2] Although the Danish /v/ is closer to an approximant than the English counterpart it is considered to be a consonant in Danish phonology (Gr nnum, 1998)

that 94 dB SPL 1 kHz calibration tone corresponded to the maximum level of the harddisk recorder.
The prompting material was played back by a Revox B226 Compact CD player over a DynAudioAcoustics BM6 loudspeaker attached to an AT-JR-32dB/10W amplifier at a clearly audible level.
The talkers were instructed to repeat what they heard including carrier sentences as outline in Table 2, F0 and nonsense syllables. They were instructed to do so at a natural level of vocal effort. The first list was presented in its entirety and followed by a short break. Subsequent lists were recorded either singly or in sequences of two or three. Test subjects were frequently offered water and breaks between lists.

| | | | | | |
|---|---|---|---|---|---|
| "Nu | bliver der | sagt: | PiTu | PiTu | PiTu" |
| now | is-being | said: | PiTu | PiTu | PiTu |

Table 2: Sample from the nonsense subcorpus: A speech prompt to be repeated/imitated by the talker

### 2.2. Read Sentences

The second set of speech material targets investigation of later processing stages and consists of whole sentences. The 20 sentences read by each talker were selected randomly from the Danish standard text corpus PAROLE (Henrichsen, 2007). Each sentence used in PiTu contained between 8 and 18 words, no sentences contained proper names, and all participating words were in a standard list of 20,000 most frequent word forms for Danish. An example is shown in Table 3

| |
|---|
| Som De/ hørte/ anklageren/ sige,<br>er der/ faste/ takster<br>for/ spirituskørsel |
| as you/ heard/ the-prosecutor/ say<br>are there/ fixed/ charges<br>for/ drunk-driving |

Table 3: Example of a Danish sentence from PAROLE

#### 2.2.1. Recording procedure

The recording setup was identical to the setup used for nonsense syllables. Talkers were given a sheet of paper with an orthographic representation of the sentences. They were then instructed to read each of the 20 sentences silently once and subsequently aloud three times in succession. Talkers were allowed breaks at any point between two sentences, and instructed to take a break halfway (after ten sentences). The 20 sentences can be seen in Table 4.
The recordings were performed in sessions interleaved with the nonsense-syllables as described in the previous section in the following way. After each talker had completed the 65 nonsense-syllables shown in Table 1, they were instructed to read sentences one through ten from Table 4. Subsequently, talkers were asked to speak the 65 nonsense-syllables again followed by sentences 11 thorough 20 from Table 1. Finally, talkers spoke the 65 nonsense-syllables for the third time.

### 2.3. Postprocessing the raw recording data

The raw recordings were then segmented in analytical units, annotating the corpus with time codes for (i) each nonsense syllable, and (ii) each syllable in the read-loud sentences. Finally, we scored each stimulus-response pair for suitability (Christiansen and Henrichsen, 2011). This turned out to be necessary, given that our group of test talkers were selected primarily among science students rather than linguistic students. Most talkers were entirely new to language tasks of this kind. Some did not pay proper attention to the phonetic fine structure of the nonsense syllables, and some had difficulties with reading the sentences properly. Even though we collected a fair amount of meta-linguistic data for each participant, we did not ask specifically for problems with dyslexia or speech disorders.

## 3. Using PiTu - the case of prosody

Each language has its own prosodic patterns and habits, and so does each individual talker, each emotional state, and each generic communication situation. Even in simple everyday conversations, all these principles interact in complex ways which must be mastered by the interlocutors since lack of prosodic control in encoding and decoding of linguistic sensitive information can be potentially disastrous. Just imagine the effects of an utterance like "I do" pronounced in the church in a less than convincing manner. The literature on prosody, huge as it may be, offers surprisingly little consensus concerning the role of prosody in spoken language communication. Does the semantic content and the pragmatic function of an utterance govern its prosodic contour? Or is prosody better described as a melodic coating applied by semantically blind rules in a simple stimulus-response circuit? Pragmaticists and functional linguists tend to take the first stance while phoneticians and speech technologists typically describe prosody in more mechanistic terms. Yet no one denies that prosody is an ever-present companion to the spoken words, and that the application processes are highly controlled. As we will argue, corpora like PiTu with a systematic variation of meaningful and nonsense utterances for the same talkers can offer new insights into the difficult issues of semantics in prosody. Using corpus data the broad and abstract questions can be complemented by simpler and more quantifiable ones such as: Which parameter is the better predictor of the prosodic contour, the identity of the talker or the semantic status of the utterance (sense vs. nonsense)? In other words, if the individual talker tends to repeat the same prosodic contour for sensical and nonsensical utterances, this would suggest that prosodic cues are personal fingerprints rather than semantic constituents (under the given experimental conditions) - and vice versa. As a demonstration, we present a simple experiment based

1. Posedamer og andre, der lever på gaden, skal have hjælp, mener et enigt Folketing.
2. Derfor skal lederne af de politiske partier påtage sig et direkte ansvar.
3. Mørket vrimler med politifolk, der afspærrer den lille gade i begge ender.
4. Men kræften fylder stadig meget i hans tilværelse.
5. Imens sidder fjorten irakiske ingeniører fordelt på fem danske asylcentre og venter.
6. Det har de gjort i næsten to år.
7. Som De hørte anklageren sige, er der faste takster for spirituskørsel.
8. Og De kan straks sige, at De appellerer dommen.
9. En regering, der fremsætter sådanne forslag, kan ikke have lønmodtagernes tillid.
10. Med ulykkeligt ansigtstræk listede danskeren stilfærdigt ind i hallen.
11. Her blev han i foyeren modtaget af snesevis af nye gratulationer.
12. Han skal i lighed med de andre danske spillere i kvalifikation for at komme ind i varmen.
13. Og da vi satte det hele sammen, fungerede det.
14. Handlingen var diskvalificerende uanset det samlede handlingsforløb og til rødt kort.
15. Indholdet var der derimod ikke meget ved.
16. I hans sidste billeder er flugtvejen næsten forsvundet.
17. Og det sjove er man troede sgu fuldt og fast på ham.
18. Han må blive boende på slottet, men får en dag besøg af kosmetiksælgeren.
19. Jeg kastede mig hurtigt frem mod den, og det lykkedes mig at få fat om halsen på den.
20. Det vil hindre mange misforståelser mellem vore to partier.

Table 4: The twenty sentences used in the PiTu corpus. The sentences are from the PAROLE coprpus

on PiTu corpus data and quantitative analysis. For the experimental design we needed a formal definition of "prosodic contour". Since we did not find a generally accepted definition of prosody in the literature - let alone a computationally applicable one - we chose to simply determine the prosodic contour of an utterance as its slow pitch variations (i.e. its variations in the supra-syllabic time domain). Of course, several other formal definitions are possible, but for our demonstrational purposes we opted for maximal tractability. Operationally speaking, we compared the prosodic contours of the utterances by comparing the global slope of their fundamental frequency (F0) graphs using linear regression. Contrary to what many linguists take for granted, fundamental frequency is not an objective property of a sound signal, and in consequence F0 resolution is not a well-defined procedure. For a prolonged full vowel, the pitch can usually be determined with little uncertainty, but for more complex mixtures of harmonic and noise components (e.g. in creaky and semi-voiced speech sounds) not even humans may have a clear sense of a fundamental frequency. For such speech signals, or parts of signals, automatic tools for pitch analysis tend to produce random results, so in order to avoid phantom values we chose a safety-first solution with cancellation of dubious data points. Using praat (Boersma, 2001) (www.fon.hum.uva.nl/praat/) for speech signal analysis, we first derived two F0 measurements for each 5 ms time slice with two independent algorithms, one for the unfiltered signal and one using Hann filtering[3]. We then switched to a 10ms segmentation of the sound files each time cell thus representing four raw F0 samplings (2x2). Only time cells meeting all three conditions below were considered as qualified, providing one valid data point each as the averaged value of the four frequency measurements contained.

1. All four values are defined
2. Each value is in [50Hz .. 500Hz]
3. Numerical range of values is less than 10Hz

The F0 estimates were converted to a semitone scale with 50 Hz as the reference and fitted to a first degree polynomial minimising the squared errors. This resulted in a slope (in semitones per second) and an intercept (an example is shown in Fig. 1) for every recorded CV-triplet and sentence.
The slope was used to answer whether the identity of the talker or the semantic status of the utterance (sense vs. nonsense) is the better predictor of the prosodic contour. We did this by calcualting a two-way anova. The results are shown in Table 5.
Since the probability that samples are drawn from the same distribution across talkers is virtually zero (0.0006), we can conclude that talker is a better predictor of F0-slope than utterance type. Moreover, the interaction between talker and utterance type is also signifficant. We interpret this as saying that the way talkers produce differences between utterances is idiosyncratic at least in terms of their F0-slope.

## 4. Final remarks

Enhancing the understanding of the relationship between early and later process of speech perception

[3]Fundamental frequencies were measured using psc script settings (a) and (b), see praat documentation for details. (a) noprogress To Pitch (ac)... 0.005 75 15 yes 0.03 0.45 0.01 0.4 0.14 600, and (b) Filter (pass Hann band)... 50 1000 100; noprogress To Pitch (ac)... 0.005 75 15 yes 0.03 0.45 0.01 0.4 0.14 300

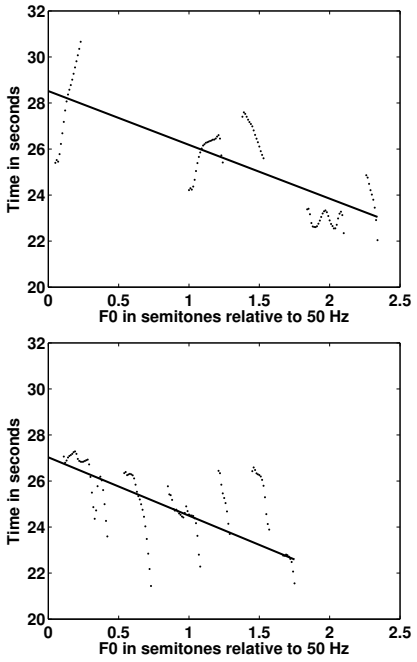


Figure 1: Example of F0 values for the Pitu material. The upper panel shows talker AH enunciating the CV-triplet of [vu:tu]. The lower panel shows AH enunciating sentence six "Det har de gjort i to år". Note the y-axis is expressed in semitones relative to 50 Hz, which corresponds to the minimum frequency considered in the present study.

| Source | Sum Sq. | Mean Sq. | F | Prob>F |
|---|---|---|---|---|
| X1 | 234.4 | 11.7191 | 2.36 | 0.0006 |
| X2 | 5.1 | 5.1465 | 1.04 | 0.309 |
| X1*X2 | 377.7 | 18.8831 | 3.8 | 0 |
| Error | 20461.3 | 4.9712 | | |
| Total | 24599.2 | | | |

Table 5: Two-way anova of F0 slope as fitted to the data (see text for details). X1 represents talkerID and X2 represents "utterance type" i.e. CV or Sentence. X1*X2 shows the interaction between X1 and X2

would facilitate progress in applied sciences such as speech recognition, speech synthesis, hearing aids, cochlear implants and telecommunication. Moreover, it would advance the theoretical basis for understanding speech perceptual processes. The speech material presented here is ideal for investigating the relationship between the perception of individual phonetic segments and whole words in sentences, thereby facilitating further investigations in line with (Bronkhorst et al., 1993; Bronkhorst et al., 2002) only with identical talkers for both nonsense-syllables and sentences. From a more linguistic point of view, PiTu and similar corpora can serve as basis for formal modelling of various aspects of speech. In this paper we discussed the case of prosody and presented a small pilot experiment. Prosody has often been overlooked by formal linguistics with its preoccupation with lexical-morphological tokens and grammar rules. For Danish, prosodic patterns have thus been studied far less extensively than other structural aspects. (Henrichsen, 2006) - building on inspirations from Nina Grønnum's work - is probably the only published computational model of Danish sentence prosody. Of course, prosodic models are essential for speech technology. Listening to a synthetic voioce with an awkward or misleading prosody can be extremely tiresome. It is therefore of interest to study the correlations between the prosodic contour predicted by formal models and the actual talker behaviour under carefully controlled conditions such as those used in the PiTu project.

Our test results are summarized in Table 5. In this small investigation, the slope of the prosodic contour is thus far better predicted by the identity of the talker than by the semantic content of the reading (sense vs. nonsense). In other words, each subject tended to reuse the same prosodic pattern, no matter what words were being said. Under the (admittedly somewhat artificial) PiTu test conditions, prosody thus seemed to serve as a personal identifier rather than as semantic markup.

## 5. References

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

A. W. Bronkhorst, A. Bosman, and G.F. Smoorenburg. 1993. A model for context effects in speech recognition. *Journal of the Acoustical Society of America*, 93:499–509.

Adelbert W. Bronkhorst, Thomas Brand, and Kirsten Wagener. 2002. Evaluation of context effects in sentence recognition. *Journal of the Acoustical Society of America*, 111(6):2874–2886.

T.U. Christiansen and S. Greenberg. 2012. Perceptual confusions among consonants, revisited. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20(1):pp. 147 – 161. ISSN: 1558-7916.

T.U. Christiansen and P.J. Henrichsen. 2011. Objective evaluation of consonant-vowel pairs produced by native speakers of danish. In *Forum Acusticum*.

H. Fletcher and R. H. Galt. 1950. The perception of speech and its relation to telephony. *Journal Of The Acoustical Society Of America*, 22(2):89–151.

H. Fletcher. 1920. The relative difficulty of interpreting spoken sounds of english. *Physical Review,*, 15:513–516.

N. Gr nnum. 1998. Illustrations of the ipa: Danish. *J. Int. Phon. Assoc.,*, vol. 28:99?105.

P.J. Henrichsen. 2006. Danish prosody, formalized. In J.Toivanen & P.J.Henrichsen (eds), editor, *Current Trends in Research on Spoken Language in the Nordic Countries*, pages 145–162. Oulu Univ. Press.

P.J. Henrichsen. 2007. The danish parole corpus - a merge of speech and writing. In J.Toivanen & P.J.Henrichsen (eds), editor, *Current Trends in Research on Spoken Language in the Nordic Countries, vol II*, pages 84–93. Oulu Univ. Press,.

F. Ingerslev, O. J. Pedersen, P. K. Møller, and J. Kristensen. 1968. New rooms for acoustic measurements at the danish technical university,. *Acustica*, 19:185–199.

1999. *The Handbook of the International Phonetic Association*. Cambridge University Press. ISBN 9780521637510.

G. A. Miller and P. E. Nicely. 1955. An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 22(2):338–352, March.