Speaking Styles

Advanced Signal Processing SE
WS 03/04

Robin Hofe

Contents

- Definition of ‘Speaking Styles‘
- Applications
- Methodology
- Featured Investigations:
  - Hyperactive Articulation
  - Fast and Slow Speech
  - Spontaneous and Read Speech
  - Speaking formants
  - ‘Fun‘ Speaking Style
  - Prosody Variation with Text Type
  - The LAIPTTS System
- References

Definition of ‘Speaking Styles‘ 1/2

In literature, there is so far no standard framework for classifying speaking styles.

Here, I choose the following definitions to specify my work:

• Speaking style is a set of properties by which you can link speech to a certain communicative situation.

• This situation is multi-dimensional:

- Content	(news, poem, dialogue, etc.)
- Speaker	(habits, personality, etc.)
- Situation	(distance, noise, room size, etc.)
- Listener	(relation to speaker, number, etc.)

Definition of ‘Speaking Styles‘ 2/2

- In the following chapters, we often have a pair of characteristics which will be compared (e.g. fast - slow). This is obviously not the best solution but it can give you clear results.
- As this is a very young research field and most featured studies are pilot projects, I can only give quantitative results.
- I have excluded emotional speaking styles.

Applications

- Tutoring systems, dialogue systems
- Medical applications, nursing robots
- Animated characters
- Talking newspapers for the blind
- Applications in ‘hands-busy-eyes-busy‘ situations
- Various applications in voice recognition

Methodology

How do we collect relevant data ?

- Production of ‘spontaneous‘ speech in the laboratory
- Do professional speakers produce representative results of ‘common‘ speech ?
- What do I examine in the signal (f0, formants, durations, etc.) ?

How will the data be processed ?

- Use of the procedures already introduced in previous talks.
- Which procedures are optimal ?

Featured Investigations

In the following chapters the structure will be:

- Authors
- Aim of the investigation
- Database
- Methods applied
- Results and discussion

- It is important to point out that these investigations were carried out in certain languages and apply without further studies to those only!

Hyperactive Articulation 1/7

Stefanie Köster: ‘Acoustic-phonetic Characteristics of Hyperarticulated Speech for Different Speaking Styles‘

- Aim: To describe the difference between normal and hyperarticulated speech. Parameters: duration, fundamental frequency, formants, formant bandwidth.
- Database: German speech, normal and hyperarticulated. Isolated words, sentences and spontaneous speech. 3 female and 3 male speakers.
- Methodology: For normal speech, the speaker had to address a person in front of him/her. For hyper-articulated speech, this person had headphones, signalling a communication obstacle.

Hyperactive Articulation 2/7

- Methodology (2): The spontaneous speech was created by simulating a train reservation system.

  Furthermore, an auditory test was performed to find perceptional differences between the two speaking styles: for each stimulus, the listeners had to decide how it applies to a pair of antonyms on a scale from 0 (e.g. slow) to 6 (e.g. fast).

Hyperactive Articulation 3/7

- Results: Significant differences can be observed between the two speaking styles:
- Duration increases for hyperarticulated speech in all 3 text types.

Figure 1: Duration increases for all text styles after eliminating all pauses.

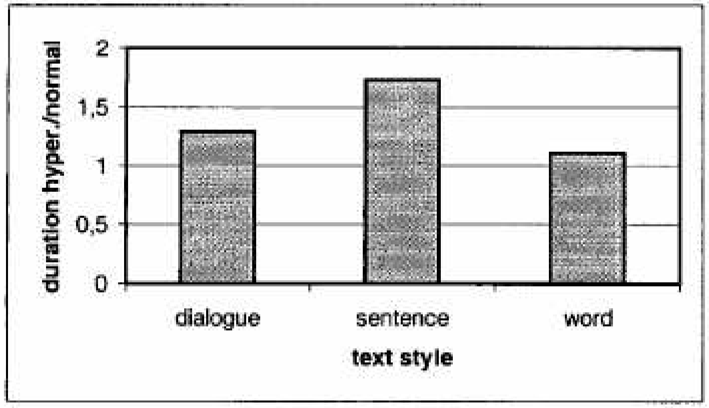
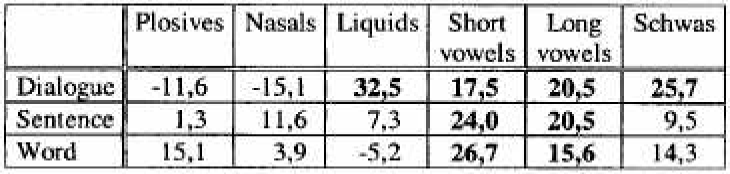


Figure 1: Ratio of duration hyperarticulated/normal speech

Table 1: Average duration change [%] for the phoneme classes.

> Vowels seem to be important.



	Plosives	Nasals	Liquids	Short vowels	Long vowels	Schwas
Dialogue	-11,6	-15,1	**32,5**	**17,5**	**20,5**	**25,7**
Sentence	1,3	11,6	7,3	**24,0**	**20,5**	9,5
Word	15,1	3,9	-5,2	**26,7**	**15,6**	14,3

Table 1 : Percentages of changes in average segmental duration

Hyperactive Articulation 4/7

Figure 2: Average change in the duration of vowels depending on their position in the sentence.

> Emphasis on the beginning and ending of a sentence seems to be important.

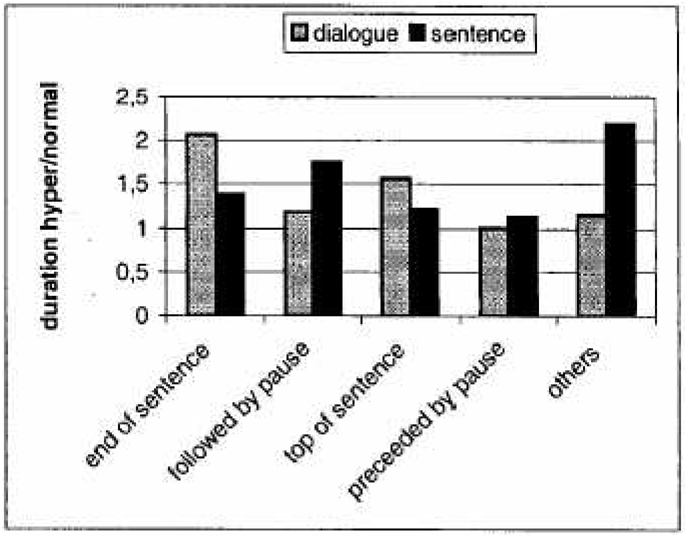


Figure 2: -Ratio of vowel duration for hyperarticulated / normal speech for different syllable positions

Hyperactive Articulation 5/7

- Fundamental Frequency: Increase visible in all three categories:

The biggest difference is visible for single words: 25%.

Spontaneous speech and sentences have 21.5% and 21.2% f0 increase.

> Great rise in the variation of f0 within spontaneous speech.

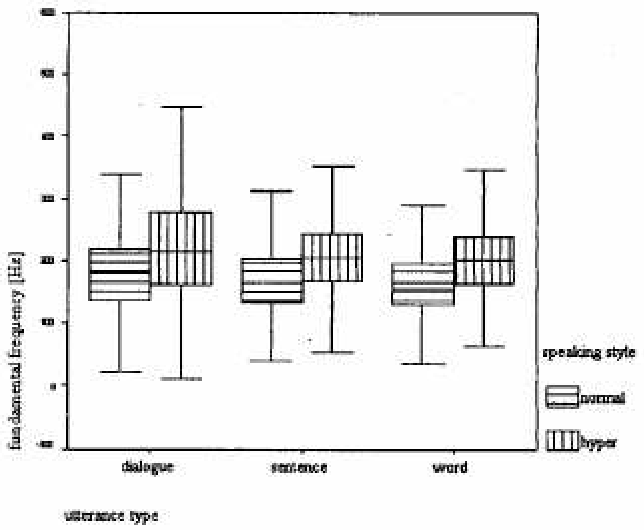


Figure 3: Average F0 in Hertz for the three utterance types

Hyperactive Articulation 6/7

- Formants and formant bandwidth: Formant bandwidths are generally lower, formant frequencies decrease:

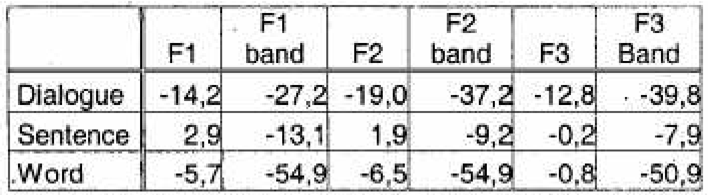
Table 2: Differences [%] for all phoneme classes.

> Sentences seem less affected.

> The tendency of decreasing formant frequencies was very strong with fricatives.

Figure 4: F1-F2 plane for long vowels, both hyperarticulated and normal speech.

> Different results have been found for English language!



	F1	F1 band	F2	F2 band	F3	F3 Band
Dialogue	-14,2	-27,2	-19,0	-37,2	-12,8	-39,8
Sentence	2,9	-13,1	1,9	-9,2	-0,2	-7,9
Word	-5,7	-54,9	-6,5	-54,9	-0,8	-50,9

Table 2: Average of formant and formant bandwidth frequency changes of all voiced phonemes in percent

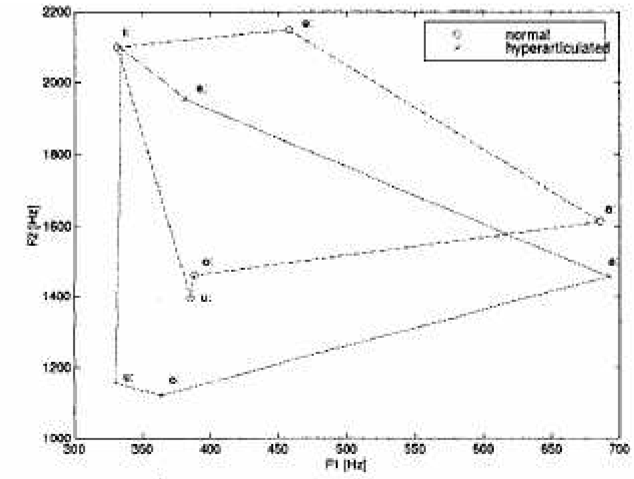


Figure 4: Formant frequencies in F1-F2 plane for long vowels

Hyperactive Articulation 7/7

- Perceptional Differences:

  > Hyperarticulated speech is perceived as fast or slow as normal speech.

  > It was judged as uncomfortable, aggressive and powerful. This might well be a problem when it comes to synthesis.

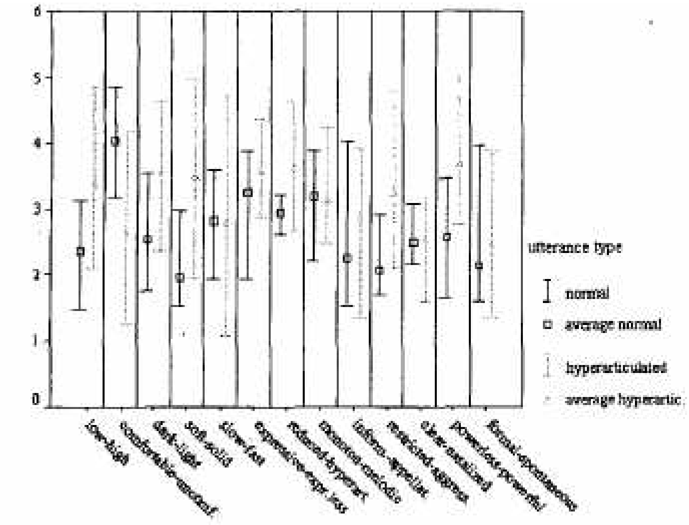


Figure 5: Minimum, maximum and average judgements for antonymes

Fast and Slow Speech 1/6

Alex Monaghan: ‘An Auditory Analysis of the Prosody of Fast and Slow Speech Styles in English, Dutch and German’

- Aim: Determining differences in accent location, boundary location and boundary strength at different speech speeds and whether there is consistency in those across languages.
- Database: The multilingual speech database, recorded in April 1999 as part of the COST 258 programme.

  Texts of the same general style (informative) but not of the same content, read in two speeds.

COST: European Cooperation in the field of Scientific and Technical Research

Fast and Slow Speech 2/6

- Database (2): Languages and speakers were:

  English / Male

  Dutch / Female

  Austrian German / Male (GermanA)

  Swiss German / Male (GermanS)

  Leipzig German / Female (GermanL)

  Bonn German / Female (GermanB)

GermanS and GermanL had the same text.

Fast and Slow Speech 3/6

- Methodology: Manual transcription of three aspects of the prosody of the recorded speech by adding diacritics to the written text: accent location, boundary location, boundary strength.

  Boundary strength was transcribed using three categories:	major pause (Utt)

  minor pause (IP)

  boundary tone, no pause (T)

Fast and Slow Speech 4/6

Results:

Table 1 gives an overview over the different text samples: they vary greatly in length and speech speed difference.

Table 1: Length in words, and duration of the six text samples

	Words	Fast	Slow	Fast/Slow
English	35	11,5s	17,0s	0,68
Dutch	75	24,0s	42,0s	0,57
GermanA	148	54,5s	73,0s	0,75
GermanB	78	28,0s	51,0s	0,55
GermanL	63	27,0s	49,0s	0,55
GermanS	63	25,5s	38,5s	0,66

- Accents:

There are never more accents in the fast version. The accents in the fast version occur only at locations, where there are accents in the slow version.

> Fast Speech can be characterized by accent deletion.

Table2: Accents transcribed and Overlap between fast and slow variety

	Fast	Slow	Overlap
English	21	21	21 (100%)
Dutch	34	43	34 (79%)
GermanA	74	78	74 (95%)
GermanB	35	42	35 (83%)
GermanL	28	41	28 (68%)
GermanS	33	36	33 (92%)

Fast and Slow Speech 5/6

- Boundaries:

As with accents, an increase of the number of boundaries with slow speech rate is visible. The lowest raise (30%, GermanA) corresponds to the lowest decrease in speaking time.

Table 3: Total number of boundaries, all categories

	Total Boundaries	
	Fast	Slow
English	8	11
Dutch	10	21
GermanA	22	29
GermanB	12	36
GermanL	7	23
GermanS	8	15

Table 4 shows a trend towards weakening of pauses with increasing speech speed.

> Fast speech seems to be characterised by a decrease or deletion of boundary strengths.

Table 4: Changes in Boundary strength from slow to fast

	-2	-1	0	+1	+2
English	2	7	2	0	0
Dutch	5	14	2	0	0
GermanA	2	15	12	0	0
GermanB	11	17	8	0	0
GermanL	8	15	0	0	0
GermanS	3	10	2	0	0

Fast and Slow Speech 6/6

Important for TTS: In slow speech, Utt and IP boundaries correspond to punctuations and commas but T boundaries are not predictable this way.

It appears that for a given text, different speakers agree on which accents to delete or boundaries to demote on a faster speech rate.

- Conclusion: For a given text and speech rate, accent location, Utt and IP boundary locations and strengths are predictable from the text.

  Faster speech rates can be constructed from slower ones.

  T boundaries remain unpredictable.

  Another investigation by Brigitte Zellner comes to the same conclusions for French.

Spontaneous and Read Speech 1/10

Marc Swerts, Eva Strangert, Mattias Heldner: ‘F0 Declination in Read-aloud and Spontaneous Speech’

- Aim: Investigation of differences in the f0 contour between spontaneous and read Swedish speech.
- Database: 2 Swedish Speech samples spoken by a male person: A read-aloud news telegram (233 words) and a spontaneous retelling of the text (252 words).
- Methodology: Estimation of the slope of declination in the f0 contour by fitting an all-points regression line to the f0 points. The boundaries between phrases and utterances were determined by the mean scores of 9 transcribers.

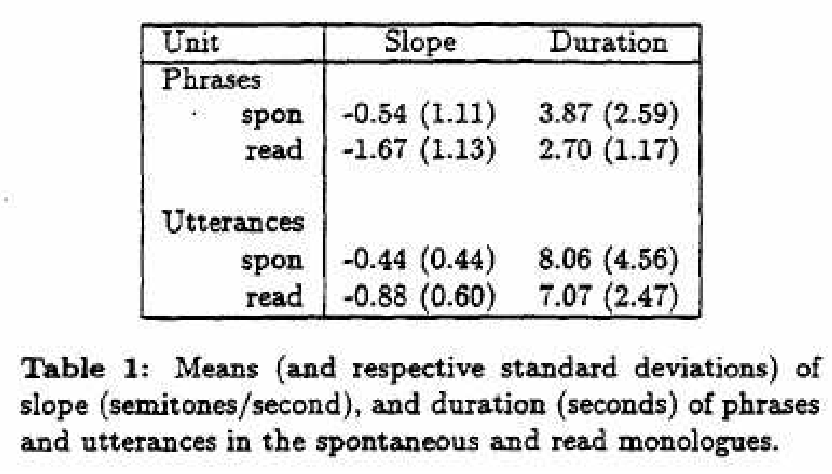
Spontaneous and Read Speech 2/10

- Results:

  > The mean f0 was higher in read-aloud (136 Hz) than spontaneous speech (107 Hz).

  > Slopes were significantly steeper in the read-aloud version.

  > Duration was shorter for read-aloud speech.



Unit	Slope	Duration
Phrases		
spon	-0.54 (1.11)	3.87 (2.59)
read	-1.67 (1.13)	2.70 (1.17)
Utterances		
spon	-0.44 (0.44)	8.06 (4.56)
read	-0.88 (0.60)	7.07 (2.47)

Table 1: Means (and respective standard deviations) of slope (semitones/second), and duration (seconds) of phrases and utterances in the spontaneous and read monologues.

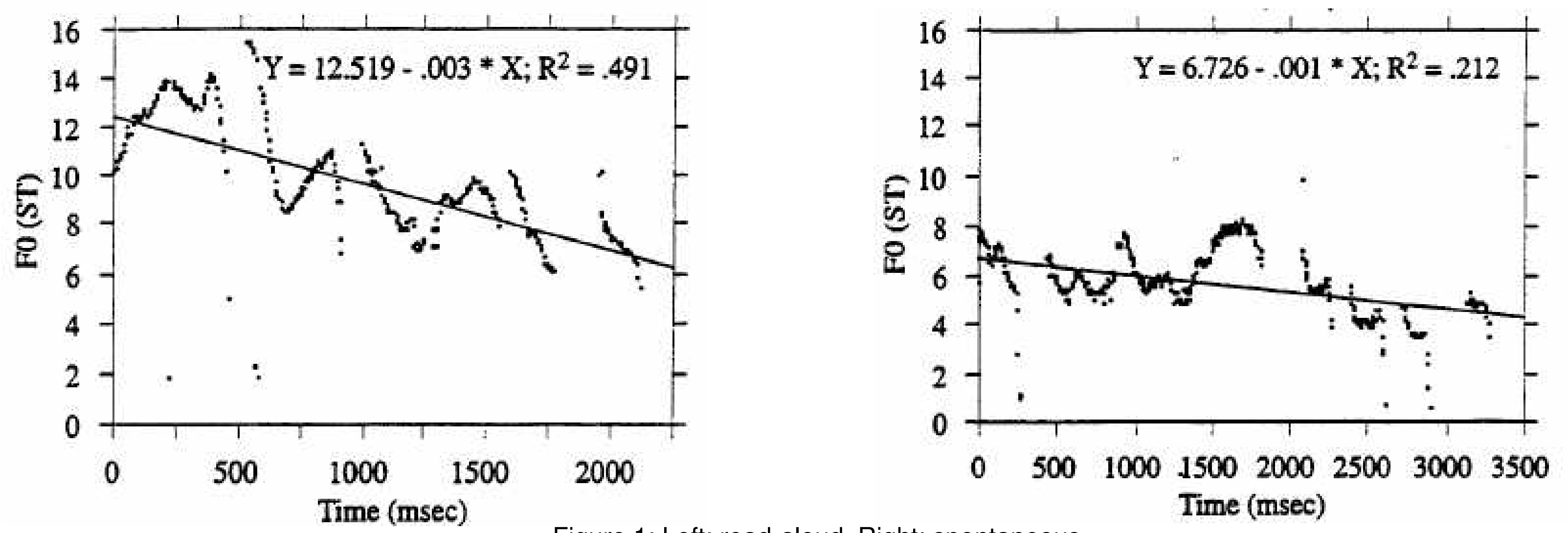


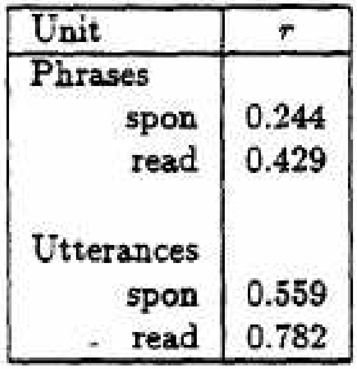
Figure 1: Left: read-aloud, Right: spontaneous

Spontaneous and Read Speech 3/10

The data from Table 1 suggests a time dependency of the f0 slope. Correlation coefficients between the length of phrases/utterances and the slope were calculated:

> Significant correlations are given.

> Speaker seems to adapt his speaking style due to the overview he has over the phrases in the near future.



Unit	r
Phrases	
spon	0.244
read	0.429
Utterances	
spon	0.559
read	0.782

Table 2: Correlation between declination and duration of phrases and utterances in the spontaneous and read monologues.

Spontaneous and Read Speech 4/10

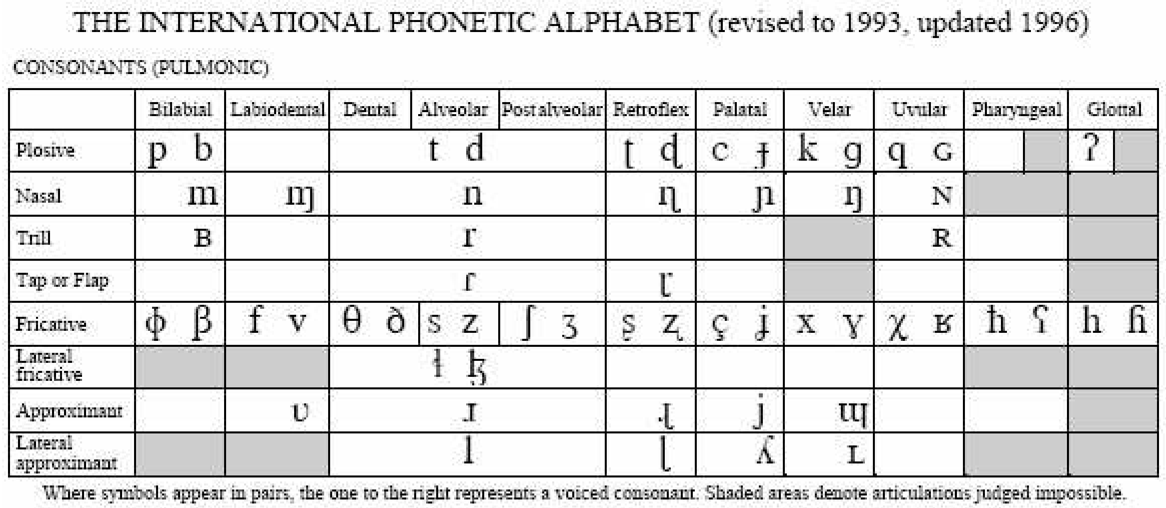
Danielle Duez: ‘Reduction and Assimilatory Processes in Conversational French Speech Implications for Speech Synthesis’

Aim: To give an overview over effects that happen in conversational French compared to ‘hyper articulated’ French.

He uses examples of widely known conversational French and analyzes them linguistically.

In the following, I will shorten or even omit his explanations and stick to the conclusions he draws from the examples.

Spontaneous and Read Speech 5/10



THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993, updated 1996)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap				ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ						▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

- Reduction and contextual assimilation:

Voiced Stops:

(1) There is a partial or complete nasalization of /b/ and /d/ proceeding or succeeding a nasal vowel.

Example: <*pendant*> /pãdã/ → /pãnã/

Spontaneous and Read Speech 6/10

(2) Weakening of /b/ into /ß/ and /w/ and of /d/ into /z/ or /l/ or its complete deletion.

These changes are assumed to result from a reduction in the magnitude of the closing gesture. Note that the place of articulation stays the same.

Consonant Sequences:

$[C_1\#C_2]$ or $[C_1C_2]$, where ‘#’ marks a syllable boundary.

Often, C_1´s were changed into another consonant or omitted. Voiced/unvoiced fricatives and plosives were devoiced/voiced due to the anticipatory effect of an unvoiced/voiced C_2.

Spontaneous and Read Speech 7/10

Examples:

<*Il m´est arrivé*> /ilmɛtaʀive/ → /imɛtaʀive/
Omission of the /l/.

<*Une espéce de*> /ynɛspɛsdə/ → /ynɛspɛzə/
Fusion of /sd/ to /z/.

- Factors limiting reduction and assimilation effects:

Segment properties:

Dentals show a higher resistance against reduction and assimilatory effects than labials.

Sonorants were the consonants that were omitted most frequently.

Spontaneous and Read Speech 8/10

Voiced occlusives are less resistant than unvoiced ones, having less articulatory force.

The more resistant segments could in turn be the ones with stronger influence on their neighbours.

Syllable structure:

Highest identification score for fricative-sonorant sequences, lowest for occlusive-occlusive.

Higher identification rate for homosyllabic consonant sequences than for heterosyllabic.

Spontaneous and Read Speech 9/10

Word Class:

Elided article or preposition <*de*> is often omitted or changed to /n/.

The <*j*> of <*je*> often changes or removes the first consonant of a following verb.

- Tendencies in reduction and assimilation:

Nasalization:

Nasalization in French tends to spread from one segment to another.

Spontaneous and Read Speech 10/10

Resistance of prominent final-phrase syllables:

There seem to be important landmarks, which attract the listeners special attention and are crucial for comprehensibility.

Speaking Formants 1/3

Gunilla C. Thunberg: ‘Balancing Spectra Between Different Speaking Styles’

- Aim: Exploring the use of spectral balance in different speaking styles: Is there a ‘speaker’s formant’?
- Database: 2 female professional actors performed the same text by first using their stage voice and then a conversational tone for close range.
- Methodology: Recordings under rehearsal-like conditions on a stage.

  Spectrum analysis was applied to the recordings.

Speaking Formants 2/3

- Results:

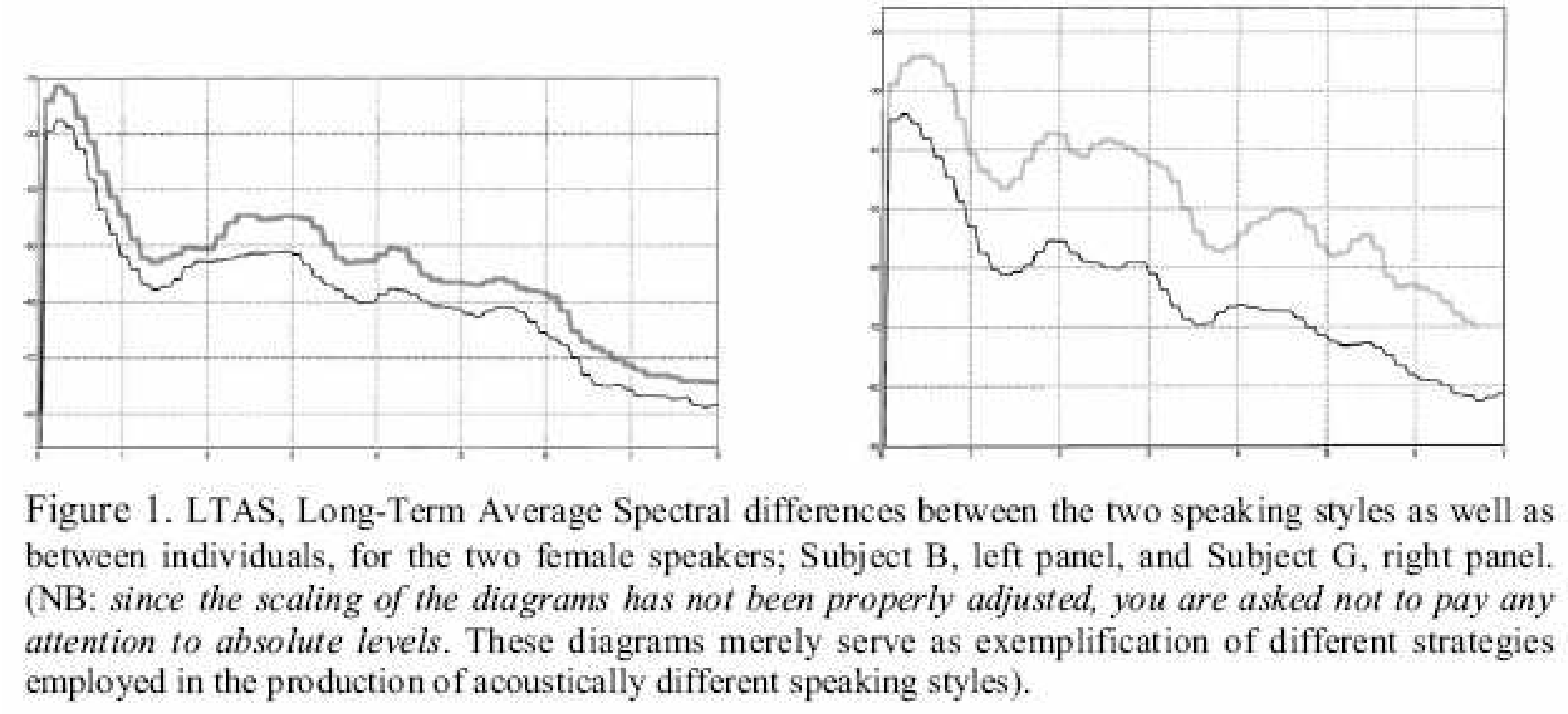


Figure 1. LTAS, Long-Term Average Spectral differences between the two speaking styles as well as between individuals, for the two female speakers; Subject B, left panel, and Subject G, right panel. (NB: *since the scaling of the diagrams has not been properly adjusted, you are asked not to pay any attention to absolute levels*. These diagrams merely serve as exemplification of different strategies employed in the production of acoustically different speaking styles).

SubjectB (left) has higher f0 values altogether.

Speaking Formants 3/3

- Discussion: SubjectB has apparently no change in the spectral contour.

  While SubjectG could be compared to an alto singer, SubjectB is more like a soprano.

  Due to their very high f0 and thereby fewer overtones, sopranos do generally not use a singer’s formant.

  SubjectG shows 3 extra peaks: at 2-3 kHz (where the singer’s formant would be) but also at 4-5 kHz and 5-6 kHz.

Further investigation will be necessary.

‘Fun‘ Speaking Style 1/6

Kjell Gustafson, David House: ‘Prosodic Parameters of a ‘Fun’ Speaking Style’

- Aim: To examine how children react to prosodic variations differing from the standard TTS for adult audiences.
- Database: 3 sentences, spoken by the Infovox 230 formant-based Swedish male voice and the Infovox 330 concatenated diphone Swedish female voice.

  Four different versions of each sentence were produced: (1) standard, (2) double duration in focused words, (3) double of maximum f0 values in focused words, (4) combination of (2) and (3).

‘Fun‘ Speaking Style 2/6

- Methodology: On a computer, an animated character was shown (an astronaut). 8 text fields next to it were linked to sound files. The subjects could activate those by clicking on them.

  They were then asked, which example they found the most natural one and which the most ‘fun’.

  Subjects were 8 children and 4 adults as a control group.

‘Fun‘ Speaking Style 3/6

- Results:

> Large manipulation of both f0 and duration were preferred for a ‘fun’ voice.

> For naturalness, children preferred higher f0 peaks.

> The f0 manipulations were found to be more ‘funny’ by both adults and children.

> Children distinguish strongly between naturalness and fun.

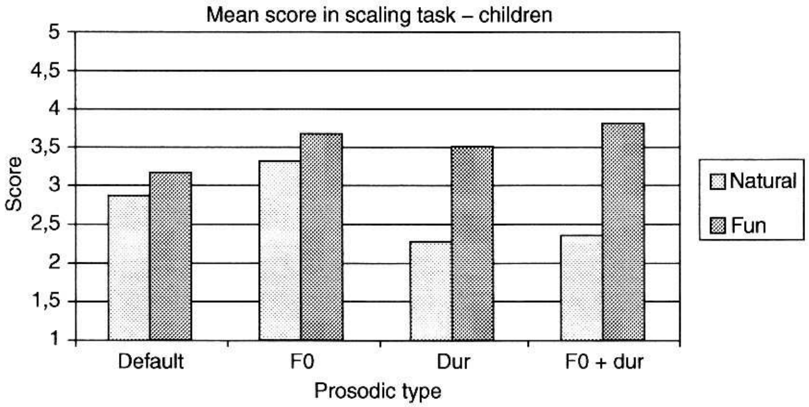


Figure 26.3 Comparison between fun and naturalness scaling – children

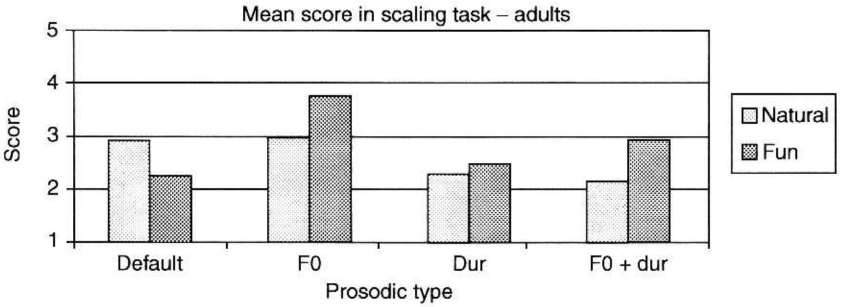


Figure 26.4 Comparison between fun and naturalness scaling – adults

‘Fun‘ Speaking Style 4/6

This figure shows the result of a ranking task for the three sentences.

> ‘Extreme’ prosody is clearly preferred in both naturalness and ‘fun’.

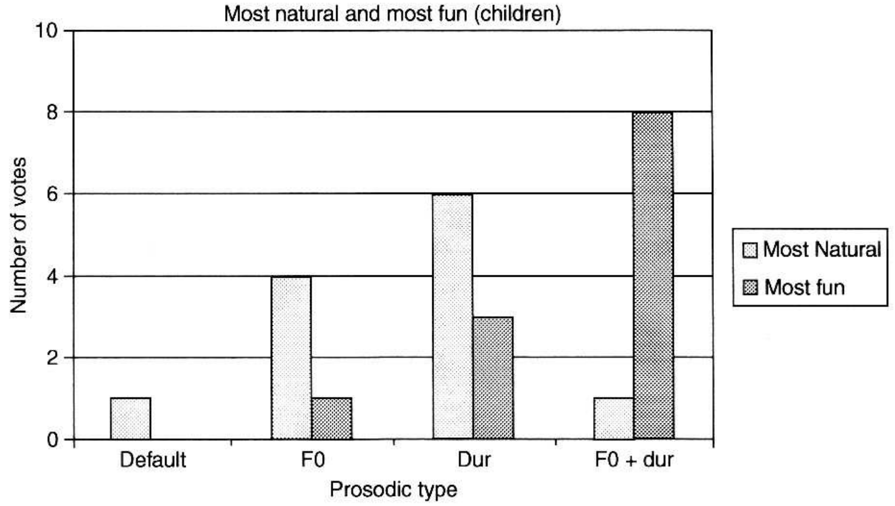


Figure 26.5 Children’s ranking test: votes by four children for different realisations of each of three sentences

The children preferred the formant synthesis over the diphone synthesis, the reason could be that it was seen as more adequate for an animated character.

For a larger character (lion, elephant), a low f0 might be preferred.

‘Fun‘ Speaking Style 5/6

- Sound examples:

The sentence presented is:

‘Idag ska jag flyga till en ANNAN planet.’

‘Today I’m going to fly to a DIFFERENT planet.’

Formant Synthesizer, default

Formant Synthesizer, f0

Formant Synthesizer, duration

Formant Synthesizer, f0+duration

Diphone Synthesizer, default

Diphone Synthesizer, f0

Diphone Synthesizer, duration

Diphone Synthesizer, f0+duration

‘Fun‘ Speaking Style 6/6

- Tasks for future studies:

Investigate ideal values for the modifications and optimum shapes of the f0 contour.

Should the modifications be restricted to focused words?

Is the effect the same with longer texts?

Prosody Variation with Text Type 1/7

‘The Variation of Prosody with Text Type’, Justin Fackrell, Halewijn Vereecken, Jean-Pierre Martens, Bert Van Coile

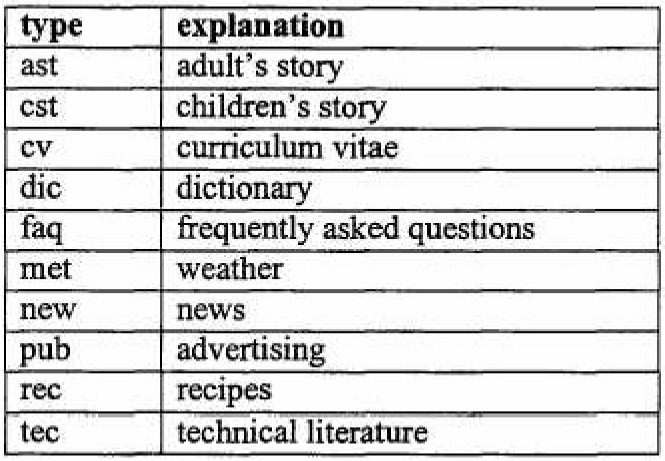
- Aim: Examination of text type dependent variations of prosody on a document lever rather than on a sentence or word level.
- Database: Texts from 10 text types in three languages (Dutch [Belgium], English [US] and French [France]) were collected from the internet and read by female native speakers. (3 hours of material)
- Methodology: Word prominence (PRM) and inter-word prosodic boundary strength (PBS) were transcribed using a 0..5 scale.

Prosody Variation with Text Type 2/7

- Methodology (2): A pitch extractor determined a pitch track for each recording.

  The grapheme-to-phoneme component of the RealSpeak^{TM} TTS engine was used to get a prototypical phonetic transcription of the texts.

  Word length in syllables and average phrase length in syllables were measured.



type	explanation
ast	adult's story
cst	children's story
cv	curriculum vitae
dic	dictionary
faq	frequently asked questions
met	weather
new	news
pub	advertising
rec	recipes
tec	technical literature

Table 4 The 10 text types used in the pilot experiment

Prosody Variation with Text Type 3/7

- Results:

Figure 1: Showing the word length variations from one text type to another.

> Children’s stories (but also adult’s stories) have the shortest, CV’s the longest words.

> Over all, English has shorter words.

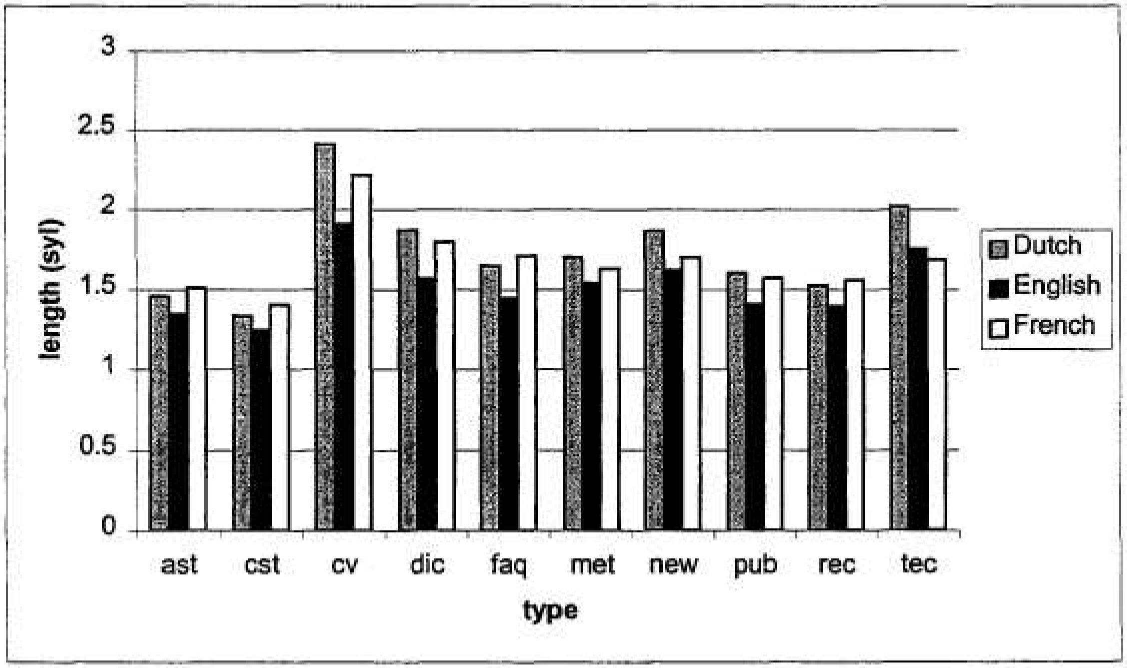


Figure 1 Variation of word length (in syllables) with text type

Prosody Variation with Text Type 4/7

Figures 2 and 3 show the speaking rate in syllables/minute and the difference from the average speed of the language per text type.

There are great differences between the text types, but also between the speakers reading the same text type.

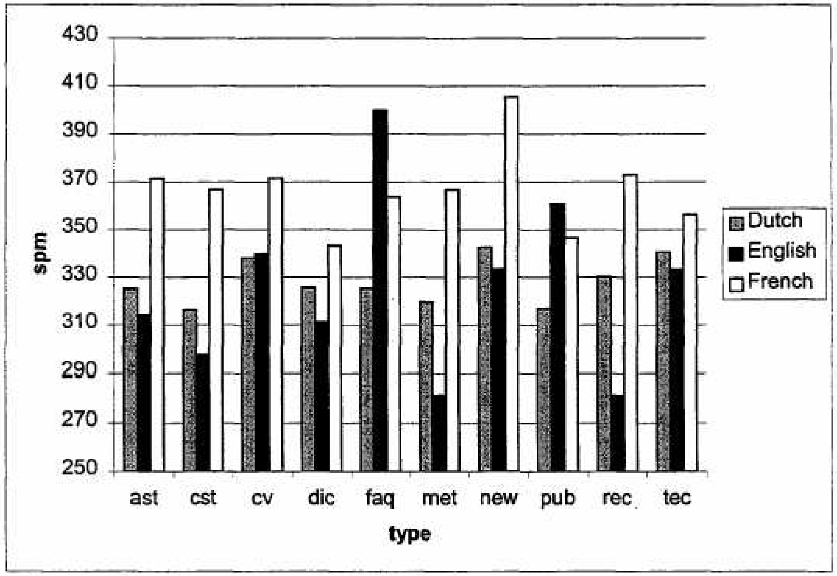


Figure 2 Variation of speaking rate (in syllables per minute) with text type

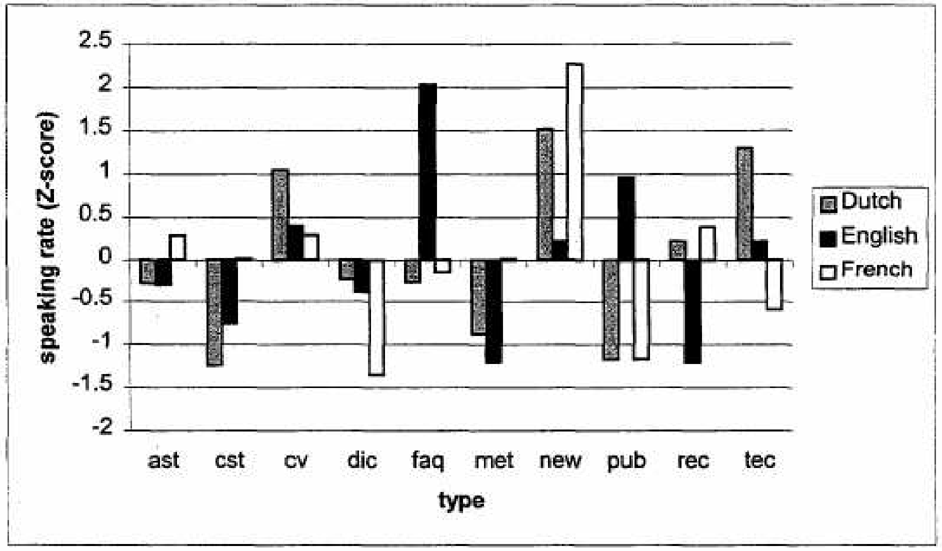


Figure 3 Variation of speaking rate (Z-score) with text type

Prosody Variation with Text Type 5/7

Figure 4 shows a scatterplot between median phrase and pause duration.

Some text types show consistency over languages.

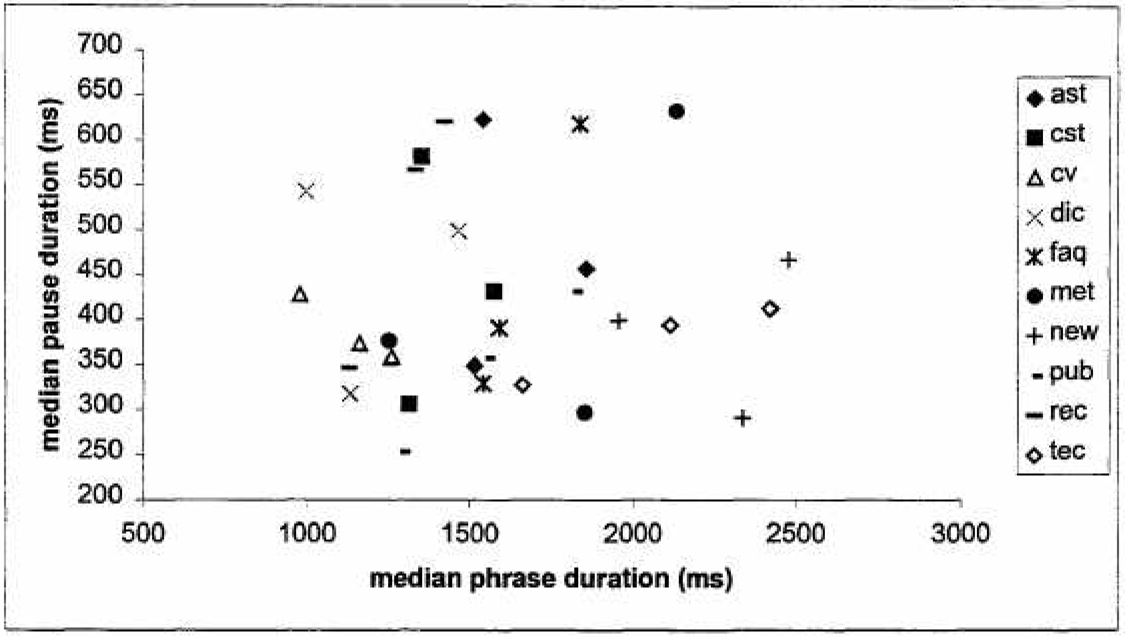


Figure 4 Scatterplot showing relation between median phrase duration and median pause duration

Prosody Variation with Text Type 6/7

Figures 5 and 6 show the variation of average pitch depending on text type and language.

There are, again, differences between languages and speakers.

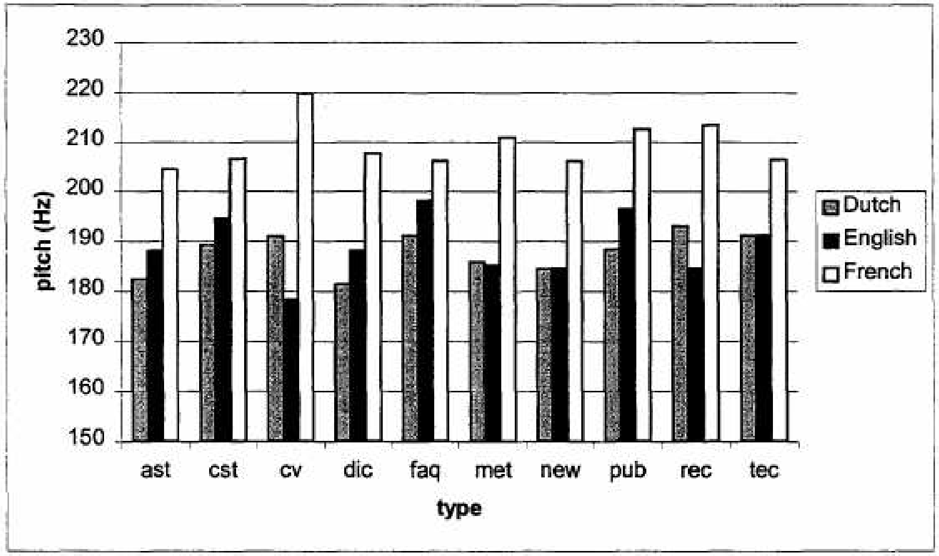


Figure 5 Variation of average pitch with text type

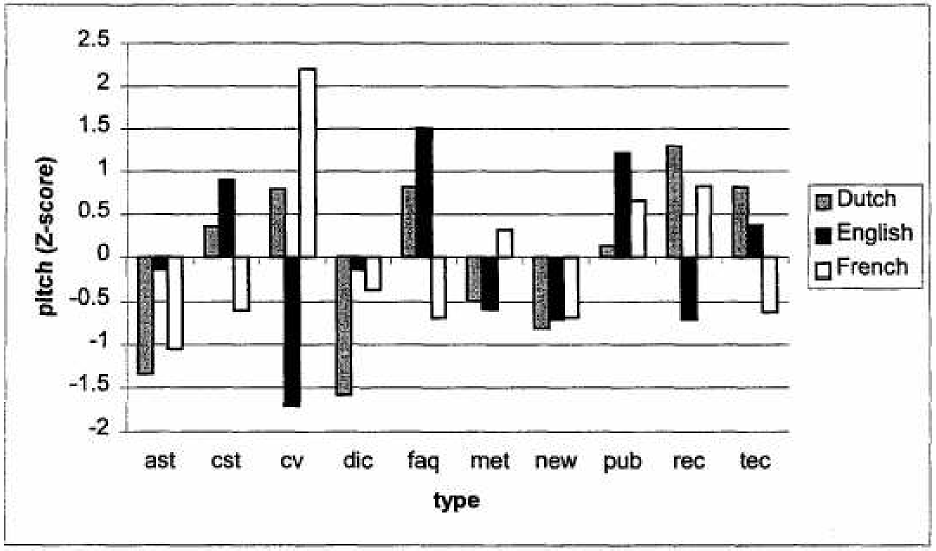


Figure 6 Variation of pitch (Z-score) with text type

Prosody Variation with Text Type 7/7

- Discussion:

  Prosody variation seems to be at least language dependant, if not speaker dependant. (Probably both)

  The strongly diverse text types seemed to be CVs and children’s stories.

The LAIPTTS System 1/5

‘A Nonlinear Rhythmic Component in Various Styles of Speech’, Brigitte Zellner Keller, Eric Keller

- Goal: Developing a dynamic model of the temporal organisation of speech. This requires a complex nonlinear dynamic model.
- Flaws of current approaches:

  Assumption of a linear relation between variables influencing speech timing.

  The variability is underestimated. Human speech gestures are never repeated exactly the same way.

  Statistical approaches often lead to a false impression of causal relation.

The LAIPTTS System 2/5

- The BioPsychoSocial Speech Timing Model:

The levels of constraints that govern speech activity in the time domain:

> (1) Bio-psychological: respiration, neurology, psycho-rhythmic tendencies.
>
> (2) Social: linguistic and socio-linguistic constraints.
>
> (3) Pragmatic: situation, feelings, cognitive tasks.

(1) is the base level, on which (2) and (3) superimpose their constraints.

The LAIPTTS System 3/5

More effort should be put into the investigation of serial temporal dependencies.

Autocorrelation methods should be applied to find dependencies of an $X_{k+1}|X_1..X_k$ type.

Applying such methods, the LAIP found regular anticorrelations for French and, to a lesser extend, for English.

The anticorrelation component manifested itself reliably within 500ms or 1 or 2 syllables.

This could have a neurophysiological and/or articulatory cause. It may even be a general human rhythmic tendency.

The LAIPTTS System 4/5

This anticorrelational effect was implemented as a parameter in the LAIPTTS Speech Mill speech synthesis system and added a ‘swingy’ effect with high values or a more ‘controlled’ effect with weak ones.

The LAIPTTS SpeechMill uses a grapheme-to-phoneme translator, the LAIP-designed prosody control system for timing and f0 and Mbrola diphone motor.

The LAIPTTS System 5/5

- Sound examples:

Welcome

Unmarked reading language

Normal speed

Slow speed

Fast speed

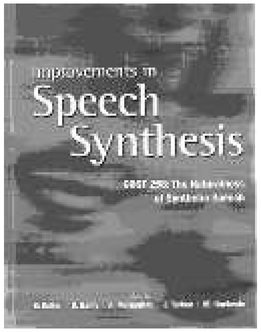
Explanation of the LAIP project

Beat Siebenhaar wohnt mit seiner Familie in der Kleinstadt Aarau, die zwischen Zürich und Bern liegt. Für die Fahrt zur Arbeit nimmt er die Bahn, die ihn mit direkten Verbindungen in zweieinhalb Stunden nach Lausanne bringt. Wenn er nicht das Familienleben mit seinen 3 Töchtern pflegt oder am Computer sitzt, fährt er manchmal seine 30-jährige goldene DS spazieren.

Resources 1/2

> ‘Improvements in Speech Synthesis’ Edited by E. Keller et al. 2002

http://www2.unil.ch/imm/cost258volume/cost258volume.htm



> ‘Acoustic-phonetic Characteristics of Hyperarticulated Speech for Different Speaking Styles‘, Stefanie Köster, Ruhr University Bochum

> ‘Temporal Structures for Fast and Slow Speech Rate’, Brigitte Zellner, LAIP, University of Lausanne

> ‘F0 Declination in Read-aloud and Spontaneous Speech’, Marc Swerts, Eva Strangert, Mattias Heldner

> ‘Balancing Spectra between Different Speaking Styles’, Gunilla C. Thunberg, Stockholm University

> ‘The Variation of Prosody with Text Type’, Justin Fackrell, Halewijn Vereecken, Jean-Pierre Martens, Bert Van Coile

Resources 2/2

> http://www2.unil.ch/imm/docs/LAIP/COST_258/cost258.htm

> http://www-oedt.kfunigraz.ac.at/hlt/

> http://ieeexplore.ieee.org/Xplore/DynWel.jsp

> http://www2.unil.ch/imm/docs/LAIP/LAIPTTS_D_SpeechMill_dl.htm

> http://www.scansoft.com/realspeak/