Information based speech transduction

PETER JUEL HENRICHSEN[1] AND THOMAS ULRICH CHRISTIANSEN[2]

[1] *Center for Computational Modelling of Language, Copenhagen Business School*

[2] *Centre for Applied Hearing Research, Technical University of Denmark, DK-2800 Lyngby, Denmark*

Modern hearing aids use a variety of advanced digital signal processing methods in order to improve speech intelligibility. These methods are based on knowledge about the acoustics outside the ear as well as psychoacoustics. We present a novel observation based on the fact that acoustic prominence is not equal to information prominence for time intervals at the syllabic and sub-syllabic levels. The idea is that speech elements with a high degree of information can be robustly identified based on basic acoustic properties. We evaluated the correlation of (information rich) content words in the DanPASS corpus with fundamental frequency (F0) and spectral tilt across four frequency bands. Our results show a correlation of certain band-level differences and the presence of content words. Similarly, but to a lesser extent, a correlation between F0 and the presence of content words was found. The principle described here has the potential to improve the “information-to-noise” ratio in hearing aids. In addition, this concept may also be applicable in automatic speech recognition systems.

INTRODUCTION

Speech communication relies on ”the receiver's recognition of the sender's intent by a given utterance”. The receiver thus examines the speech stream closely for linguistic content. The present study investigates the co-variance of concentrated linguistic information and basic acoustic properties. The goal is to identify, and ultimately predict, time intervals particularly important for speech intelligibility. Such predications potentially play a crucial role for enhancement of intelligibility in speech transducers such as hearing aids. We coin this concept Information based Speech Transduction.

BACKGROUND

Scientific language description has traditionally been formulated with reference to a stratified model of analytical levels, exemplified in Table 1. The following description will discuss the levels most relevant for the purpose at hand.

(Pragmatics)	Language function in real-world contexts
Semantics	Meaning content of language segments
Syntax	Combinatorial properties of language segments
Morphology	Segmentation of sound representations
Phonology	Abstract sound representation of speech streams
Phonetics	Perceived language sounds
(Psychoacoustics)	Acoustically determined language sounds

Table 1: The strata at the outskirts - pragmatics and psychoacoustics in parentheses - are usually not considered as parts of the linguistic syllabus. The level of abstraction increases with shading from bottom to top, where the bottom strata are closer to the acoustic signal

Linguistic analysis - a crash-course

Consider the sentence "It's terribly hot in here" (H.C. Andersen, from Clumsy Hans) as uttered by a human talker. A traditional linguistic analysis of this sentence would take its beginning at the phonetic level, producing an analysis along these lines as shown in Table 2.

It's	terribly	hot	in	here
ih t s	t eh r ax b l iy	hh aa t	ih n	hh ih r

Table 2: Phonetic rendering of sample sentence as annotated with CMU phonetics, Black et al. (2007)

The next step would be to relate the *phones* (i.e. the phonetic sound segments) to the corresponding *phonemes* (the mental representations of the sounds). This exercise is called a phonological transcription. At this stage, the perceived phonetic syllable [ih t s] is interpreted as the articulatory realization of four segments present in the mental lexicon, corresponding to the word string "it is" or /ih t ih z/ in terms of phonemes. Observe the bracketing conventions [...] and /.../, used for phonetic and phonological renderings, respectively. Since the details of phonological rendering are somewhat disputed (as analyses involving postulates of mental representations tend to be), we will leave it at that.

At the morphological level, the meaning-bearing segments of the string are identified, in casu "it" + "'s" + "terrib" + "-ly" + "hot" + "in" + "here". From this stage, the analysis gradually abstracts away from the auditory signal, focusing instead on the interpretation of the speech sounds rather than the sounds themselves. Morphological units, or *morphemes*, thus correspond to words (e.g. "hot", "here") or subparts of words (e.g. "-ly"). In the sample utterance, the suffix "-ly" marks "terrib*ly*" as an adverb, as opposed to other forms derived from the stem "terror" such as "terrib*le*" (adjective) and "terror*ize*" (verb). The set of analytical labels used for morphological annotation is known as *parts-of-speech (PoS).* We return to PoS shortly. Turning to the syntactic level, "it" is analysed as a sentence *subject*, "'s" as the *main verb*, "terribly hot" as a *predicate*, and "in here" as a prepositional *modifier*.

The lingustic analysis is completed by a semantic interpretation identifying the time and place of the scene, the thematic roles of the talker and the addressee, etc. A wide range of semantic formalisms could be engaged here to convey the information that the place-of-utterance is "here", the time is "now", the informational type is "description" (as opposed to e.g. "question", "answer", "command", or "feed-back"), the locutionary force is "affirmative" (not "negative", "interrogative", or "unspecified"), and so forth.

Several linguistic research traditions have, of course, revised the basic taxonomic hierarchy (Table 1) in various ways; however, the overall layout with its profound emphasis on stratification and discrete layers of description separated by (in principle) well-described interfaces has been shared by almost all linguistic schools, be they formal, generative, functional, psycho-linguistic, socio-linguistic, or - a fortiriori - computational (e.g. Dik (1997) and Jurafsky et al. (2009)).

A TRACTABLE MODEL OF LINGUISTIC CONTENT

As should be clear by now, the 'linguistic content' of an utterance is by no means a simple or well-delimited property, but rather a pointer to a structured description of almost fractal complexity. A computational model of linguistic content thus has to involve a massive simplification, yet without losing the descriptive accuracy and reproducibility of modern linguistics.

We suggest a point-of-departure at the morphological level. This level is strategically situated in the centre of the linguistic hierarchy, with the auditory signal still in sight and thus accommodating mappings from sound-related to content-related information.

Lexeme	**Phonetic form**	**Part-of-speech (PoS)**
It	[ih t]	pronoun
's	[s]	auxiliary verb
terribly	[t eh r ax b l iy]	adverb
hot	[hh aa t]	adjective
in	[ih n]	preposition
here	[hh ih r]	pronoun

Table 3: Morpho-phonetic mapping

Table 3 presents a simplified PoS-analysis of the sample utterance ignoring inflexional details. In a standard morphological description, words in languages like

English and Danish are routinely divided into two major groups according to their grammatical role as shown in Table 4 (e.g. Klammer et al. (2009)).

Main PoS category	**Sub PoS category**	**Grammatical role**	**Example**
Noun	(unpecified)	Content word	"Joe", "horse"
Adjective	(unpecified)	Content word	"hot", "blue"
Adverb	(unpecified)	Content word	"terribly", "now"
Verb	Content verb	Content word	"eats", "slept"
Verb	Auxiliary verb	Function word	"is", "could"
Pronoun	(unpecified)	Function word	"it", "here"
Conjunction	(unpecified)	Function word	"and", "either"
Preposition	(unpecified)	(Context dependent)	"in", "below"
Interjection	(unpecified)	(Context dependent)	"yes", "oops"

Table 4: Content words and function words (subcategories are only specified where necessary for the grammatical role assignment)

The categories of content words are not fixed in size, they keep including new and excluding old lexical elements over time. In contrast, the categories of function words are small in cardinality and very rarely accept new members. Rather than carry meaning by themselves, the function words establish the relations between the content words (c.f. “wife hit husband” and “wife hit *by* husband”).

In the experimental design presented below, the dichotomy of content words and function words plays a key role. To be more specific, we wish to study the correlation between the acoustic features of speech elements and their linguistic content.

THE SPEECH MATERIAL

We used the Danish Phonetically Annotated Spontaneous Speech (DanPASS) Grønnum (2009), Henrichsen (2011), Uneson and Henrichsen (2011) for the experiment at hand. More specifically, we used the monologue part of DanPASS consisting of 18 native talkers of Danish describing a network of coloured geometrical shapes.

Marking up the DanPASS corpus for linguistic content

The corpus includes hand-tagged markup for morphology as shown in Table 5. Words that are neither content nor functions words are excluded from our investigation as their grammatical role assignment cannot be determined based on their PoS alone.

We now have an effective, reproducible, and semantically sensitive markup procedure for linguistic content. Speech segments marked as content words can be considered as relatively content rich as opposed to those marked as function word.

ACOUSTIC ANALYSIS AND RESULTS

Basic acoustic parameters

The speech signal is characterised by five parameters (extracted with Praat, Boersma (2001)) each of which is computed in 5 ms frames with non-overlapping windows. These five parameters are: 1) the fundamental frequency (F0), and 2-5) sound pressure level in four contiguous frequency bands with corner frequencies 150, 803, 1358, 2212, and 3525 Hz (B1-B4). The four highest corner frequencies correspond to ERB numbers 14, 18, 22 and 26 respectively (see Moore 2003).

Analysis of F0

The results from F0-analysis for a single talker are shown in Table 5. Content words tend to have higher F0 than function words for this talker. We speculate that information richness co-varies with F0 in this material and perhaps for spoken Danish in general.

Description (Part of Speech)	**F0 (Hz)**	**Duration (ms)**
Adjective	128	7,350
Content adverb	128	4,525
Content verbs	127	1,460
Preposition	127	3,000
Noun	124	10,415
Interjection	124	375
Pronoun (interrogative)	123	100
Pronoun (demonstrative)	121	175
Conjunction	115	770
Determiner	114	4,170
Pronoun (personal)	113	1,160
Auxiliary verb	112	940

Table 5: The leftmost column shows the part of speech (PoS) for a single talker. Rows in dark grey indicate function words, rows in white indicate content words, and rows in light grey indicate ”undefined”, i.e., the intermediate category. The middle column shows F0 averaged across all words in a given PoS (calculation details given in the text). The rightmost column lists the total duration of all words in a given PoS. PoS are sorted by falling F0

The average F0 for each of the 18 talkers in DanPASS is shown in Fig. 1. F0 is consistently higher for content words as compared to function words for all 18 talkers. This indicates that F0 is a robust predictor of information content across male and female talkers. Preliminary statistical tests indicate that the different is indeed significant.

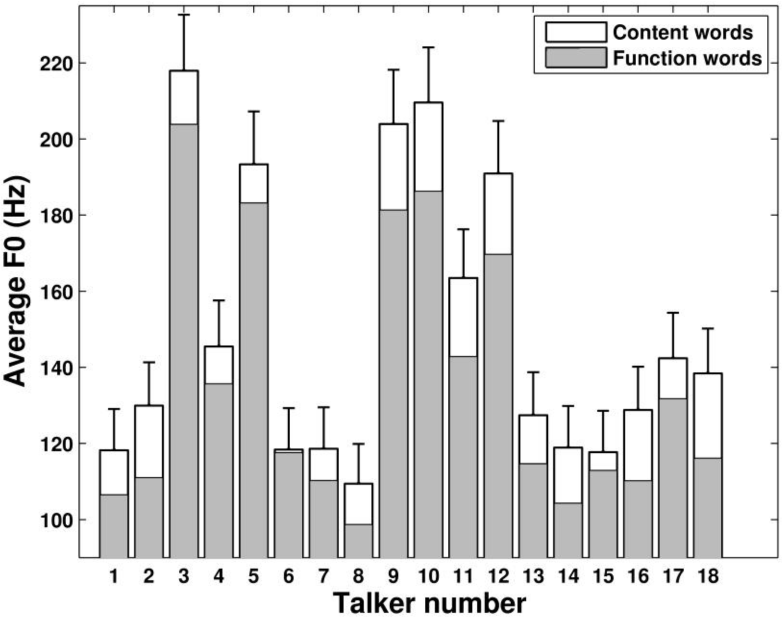

Fig. 1: Average F0 for content words (white) versus function words (superimposed in grey) as defined in Table 5 for 18 talkers. Error bars indicate one standard deviation above average F0 for content words. The 18 talkers included both females and males.

Analysis of band-level differences

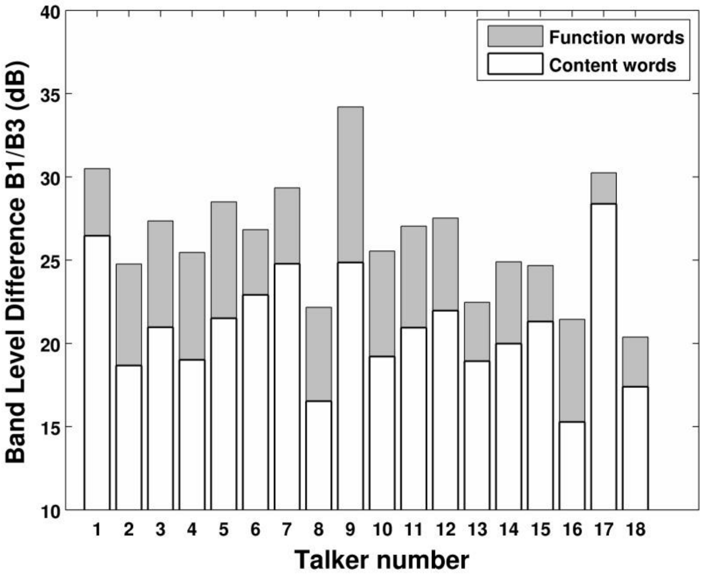

Fig. 2: Average level difference between B1 (150 to 803 Hz) and B3 (1358 to 2212 Hz). Band level differences for function words are shown in grey and band level differences for content words are superimposed in white. The same talkers as shown in Fig. 1 were used.

Fig. 2 shows the band level differences between B1 and B3 and that they are consistently lower for content words as compared to function words for all 18 talkers. This indicates that this band level difference is a robust predictor of information content across male and female talkers.

	B1	B2	B3	B4
B1		0	0	0
B2	18		7	8
B3	18	11		12
B4	18	10	6	

Table 6: The number of talkers with higher band level differences for either function words or content words for a given band combination. Cells above the diagonal (in white) show the number of talkers exhibiting higher band level differences for content words. Cells below the diagonal (in light grey) show the number of talkers exhibiting higher band level differences for function words. For example the level difference between B2 and B3 was higher for function words than content words for 11 talkers.

Table 6 shows that certain band level differences co-vary with function words while others do not. This indicates that these band level differences are robust predictors of information content across individual talkers.

CONCLUSIONS AND NEXT STEPS

Talkers seem to use simple acoustic cues to encode specific parts of their speech as particularly information rich. It may not be surprising in itself that the talker helps the listener by marking important words acoustically. What we do find surprising is, however, the lack of technological utilization. We have not been able to identify any reports of speech transducing technology (be it telecommunication, hearing aids, or ASR) exploiting the direct relation between simple physical properties and highly abstract linguistic content.

The authors are preparing a follow-up to the reported experiment using its results in an algorithm for prediction of information richness with extremely short time delay. The algorithm will be used for modulation of speech materials masking out low-content and high-content parts of the signal respectively. The manipulated signals will then be scored for intelligibility in a perception experiment. Hopefully, the results will pave the way for a new technology with a *flair* for speech.

REFERENCES

Black, A. W. and K. A. Lenzo (**2007**), “Building Synthetic Voices, Carnegie Mellon University”, www.festvox.org/, www.cs.cmu.edu/~lenzo/.

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

Dik, S. C. (**1997**), “The Theory of Functional Grammar, Part 1: The Structure of the Clause”, 2nd ed., Berlin: Mouton de Gruyter.

Grønnum, N. (**2009**), “A Danish phonetically annotated spontaneous speech corpus (DanPASS), Speech Communication 51, 594–603.

Henrichsen, P. J. (**2011**), **“**Fishing in a speech stream, angling for a lexicon”, Proceedings of 18th Nordic conference of computational linguistics NODALIDA, Pedersen, B.S., Nespore, G. and Skadina, I. (Eds.), pp. 90–97.

Jurafsky, D. and Martin, J. H. (**2009**), “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”, 2nd edition. Prentice-Hall.

Klammer, T., Schulz, M.R. and Volpe, A.D. (**2009**), ”Analyzing English Grammar”, 6th edition, Longman.

Moore, B. J. C. M. (**2003**), “An introduction to the psychology of Hearing”, 5th edition, Academic Press, pp. 72-75.

Uneson, M. and P. J. Henrichsen (**2011**), ”Expanding a Corpus of Closed-World Descriptions by Semantic Unit Selection”, accepted for the Proceedings of Computational Linguistic Applications, Warsaw, October 2011