BNAM2012
Joint Baltic-Nordic Acoustics Meeting
June 18th - 20th 2012 Odense, Denmark

Speech transduction based on linguistic content

Thomas Ulrich Christiansen
Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads, Building 352, DK-2800, Lyngby, Denmark, tuc@elektro.dtu.dk

Peter Juel Henrichsen
Center for Computational Modelling of Language,(CMOL), Department of International Language Studies and Computational Linguistics, Copenhagen Business School, Copenhagen, Denmark, pjh.isv@cbs.dk

Digital 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. This paper investigates the recent observation that speech elements with a high degree of information can be robustly identified based on basic acoustic properties, i.e., function words have greater spectral tilt than content words for each of the 18 Danish talkers investigated. In this paper we examine these spectral tilt differences as a function of time based on a speech material six times the duration of previous investigations. Our results show that the correlation of spectral tilt with information content is relatively constant across time, even if averaged across talkers. This indicates that it is possible to devise a robust method for estimating information density in the speech signal based on computationally simple short-term band-level differences. The principle described here has the potential to improve speech transduction in hearing aids and cochlear implants. In addition, the concept of information-based speech transduction may also be applicable in automatic speech recognition systems.

1 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 understanding speech. Such predications potentially play a crucial role for enhancing speech understanding in hearing aids and other hearing related technologies. We coin this concept Information based Speech Transduction.

2 Methods

2.1 Speech material

We used the Danish Phonetically Annotated Spontaneous Speech (DanPASS) [1,2,3] for the experiment at hand. More specifically, we used the monologue part of DanPASS consisting of 18 native talkers of Danish. In addition to the description of a network of coloured geometrical shapes, used in [4], we used the map task and house-building subsections of the corpus. The resulting duration of the speech material used here was six times that employed in [4]. The corpus includes hand-tagged markup for morphology as shown in Table 1. Each word in the DanPASS corpus is manually marked up for PoS (part-of-speech) according to the PAROLE convention as described in [5].

Following standard lexicographic practices, we divided the PoS taxonomy into Content and Function words[1] as shown in Table 1. Words outside of the Content and Function categories are excluded from our investigation, as their semantic status cannot be determined based on the PoS classification alone.

Table 1: The conventions used for classification of content words versus function words

Part-of-speech	Classification	Example (Danish, English)
Adjective	Content word	gul, yellow
Content adverb	Content word	ned, down
Content verb	Content word	gentage(r), repeat(s)
Preposition	Not used	over, above
Noun	Content word	trekant, triangle
Interjection	Not used	undskyld, sorry
Pronoun (interrogative)	Function word	hvor, where
Pronoun (demonstrative)	Function word	dèr, there
Conjunction	Function word	men, but
Determiner	Function word	en, a
Personal pronoun	Function word	den, it
Other pronoun	Function word	som, that
Auxiliary verb	Function word	er, is
Particle	Function word	at, to (infinitive)

Content words are relatively content rich specifying e.g. actions, mental and physical properties, and states, as opposed to function words contributing more indirectly to the meaning of the clause e.g. by specifying the relations between the other words. Mapping the PoS information onto the {Content, Function} domain thus provides us with an effective, reproducible, and semantically sensitive measure of linguistic content.

2.2 Acoustic analysis

The spectral tilt is characterised by four band-levels each of which is computed in 5 ms frames with non-overlapping windows. These four bands are contiguous and bands have 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 [6]).

3 Results

The average band level difference was reported in [4] to be greatest for B1/B3 in comparison to other combinations of bands. These differences are shown in Figure 1. The average band level differences are calculated as the mean difference of the dB values in 5 ms intervals, i.e., not as the average physical band level differences.

[1] The dichotomy of Content and Function words are also known under other names, e.g. open-class/closed-class items, or categorematic/syncategorematic types

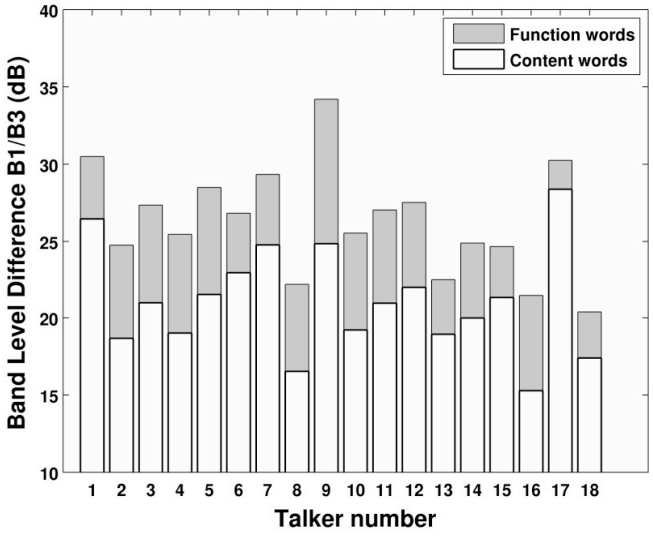

Figure 1: Average level difference between B1 (150 to 803 Hz) and B3 (1358 to 2212 Hz). Band level differences for function words are shown in grey and band level differences for content words are superimposed in white.

Figure 2 shows the band level differences for content and function words as a function of time from the beginning of the word for all combinations of bands involving B1. It shows the average band level difference is almost uniformly distributed in time, albeit function words show larger variation. This larger variation, particularly towards the end of words, can be explained by the fact that function words are shorter and fewer than content words, and thus later averages are based on fewer observations.

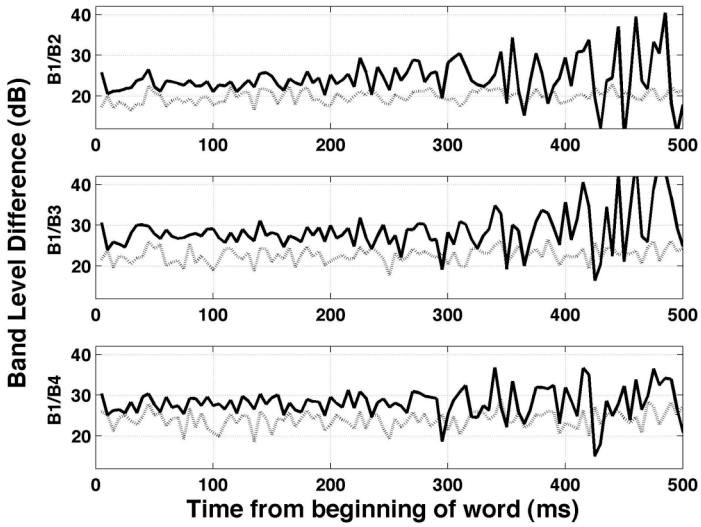

Figure 2: Average level difference as function of time. The grey curve shows the band level difference for content words and the black curve shows band level difference for function words. Top panel shows the band level difference between B1 and B2, middle panel between B1 and B3, and bottom panel between B1 and B4. Only the first 500 ms of words are shown, since function words of longer duration are rare.

4 Discussion and conclusion

The data presented here demonstrates that spectral tilt contrasts between function and content words exist for the DanPASS speech material. These spectral tilt contrasts are evident both when band level differences are averaged across time and when averaged across talkers. Although the band level differences are arguably small, it is nevertheless remarkable, that it is possible to demonstrate them using almost arbitrary (albeit contiguous) pass-bands, and arbitrary combinations hereof. We speculate that this effect would be even greater with optimal pass-bands and combinations of these. Moreover, the definition of content and function words is at present very coarse, based as it is on computational tractability rather than cognitive isomorphism. We argue that a more psychologically informed definition will allow us to see even greater spectral tilt differences between the two groups of words.

Several questions arise already from the present study. Do the observations presented here generalise to other languages and other talking situations? Is there a correlate in synthetic languages (e.g. Finnish, Inuit), which do not have the same distinction between function and content words, as do the Indo-European languages? How do individual talker characteristics and talking styles correlate with spectral tilt differences between content and function words? Is it possible to exploit such observation in hearing related technology? The later of these questions points to the potential application of this line of research.

References

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

[2] P.J. Henrichsen, P. J., 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.), 2011, 90–97.

[3] M. Uneson and P. J. Henrichsen, Expanding a Corpus of Closed-World Descriptions by Semantic Unit Selection, *Proceedings of Computational Linguistic Applications*, Warsaw, 2011, accepted.

[4] P.J. Henrichsen and T.U. Christiansen, Information based speech transduction, *Proceedings of ISAAR*: *Speech perception and auditory disorders*, 3rd International Symposium on Auditory and Audiological Research, Nyborg, 2011, in press.

[5] T. Bilgram and B. Keson, The Construction of a Tagged Danish Corpus, *Proceedings of NODALIDA*, 11th Nordic conference on computational linguistics, Copenhagen, 1998.

[6] B. J. C. M. Moore, *An introduction to the psychology of Hearing*, 5th edition, Academic Press, 2003, 72–75.