Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 4pSCb: Production and Perception I: Beyond the Speech Segment (Poster Session)

4pSCb14. An examination of the articulatory characteristics of prominence in function and content words using real-time magnetic resonance imaging

Zhaojun Yang*, Vikram Ramanarayanan, Dani Dyrd and Shrikanth Narayanan

***Corresponding author's address: University of Southern California, Los Angeles, CA 90007, zhaojuny@usc.edu**

We examine the functional coupling between articulatory characteristics of prominence, such as articulator speed, and its acoustic characteristics, such as F0 and acoustic energy, for content words (nouns) and function words (e.g., articles, prepositions and conjunctions) using real-time magnetic resonance imaging data. We use Granger causality ideas to test the degree and direction of causal influence between the chosen articulatory and acoustic measures for function and content words. We further apply functional canonical correlation analysis to these measures to understand the covariant behavioral modes of the articulatory and acoustic measures. After controlling for word duration, we observe that articulatory speed generally has a significant causal influence on F0, especially for longer content words, but observe no such effect in the opposite direction. Notably, we do not observe this effect for function words in most cases. We further observe a tighter coupling of canonical weight functions of articulatory speed and acoustic prominence characteristics for the content words as compared to the function words considered. These observations provide support for the hypothesis that prominence realized during content words may result from a close coupling between articulatory and acoustic characteristics such as articulatory speed and F0, with suggestions of a directional causal relationship.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4800676]
Received 21 Jan 2013; published 2 Jun 2013

INTRODUCTION

Speech production is realized through a composition of segmental articulations in space and time by the articulatory system that includes jaw, lips, tongue and larynx. To support human communication, speech is suited with an organized prosodic structure which is encapsulated into the articulatory as well as the acoustic signal [1]. Articulatory correlates of prosodic structure have been increasingly studied in recent research work, e.g., the effects of prosody structure or focal accent of syllables in a word [2, 3]. However, despite a growing body of work on identifying prosodic signatures at the articulatory and acoustic levels, literature on the *coordinative* spatiotemporal dynamics of articulation and acoustic of prosody remains relatively limited.

In this work, we analyze the relationships between articulatory and acoustic dynamics related to different prosodic characteristics reflected in content and function words. Specifically, we focus on addressing two questions: (1) Do we observe different causal relationships between the articulatory and acoustic correlates of prosody in function and content words, and further, (2) can we quantify the relation between the temporal dynamics of articulatory and acoustic correlates in these two cases?

Recent advances in real-time magnetic resonance imaging (MRI) enable us to capture the movement of the vocal tract in conjunction with the synchronzied acoustic signal during speech production [4], which provides an ideal technique to investigate the coordination between the articulatory and acoustic dynamics.

ANALYSIS METHODOLGY

Data Corpus and Feature Extraction

We analyzed the *read* speech of 2 native speakers of American English who spoke 460 sentences while being imaged in an MRI scanner [5]. Spoken responses and MRI videos of vocal tract articulation were recorded and time-synchronized with the audio [6]. We reconstructed the data using a sliding window technique at a rate of approximately 23 frames per second. Since MRI scanners generate a lot of noise, the recorded audio was post-processed using a custom noise-cancellation algorithm [6] before use (Further details, and sample MRI movies can be found at http://sail.usc.edu/span). We then automatically extracted the air-tissue boundary of the articulatory structures in each frame of video [7].

We obtained word-level alignments of the corpus using SailAlign, a HTK-based phonetic alignment tool [8]. We then used the Stanford Parser [9] to parse all words in the corpus and assigned part-of-speech tags to them. Based on these assignments, each word was classified as either being a content word (nouns) or a function word (e.g., articles, prepositions and conjunctions). In order to remove the confound of word duration in content and function words, we classify all words into three duration categories based on their duration distributions, $0 \sim 0.25s$ ($\mathbb{C}_1$), $0.25 \sim 0.5s$ ($\mathbb{C}_2$) and $0.5 \sim 1s$ ($\mathbb{C}_3$).

For the interval of each word, we extracted the pitch (F0) and short-term energy (E) from the speech signal as the acoustic features. We further applied a median filter to remove the spurious spikes in the pitch and energy estimates. Note that we do not consider unvoiced and silent portions in the sentence in our analysis. We use a 'gradient frame energy' (GFE) measure as the articulatory feature to capture articulator speed in content and function words [10].

Statistical Analysis

In this work, we use Granger causality (G causality) to test the degree and direction of causal influence between the chosen articulatory and acoustic measures for function and content

words. G causality is a statistical technique for analyzing causality of time series based on linear prediction [11]. Given two independent variables X and Y, X "Granger-causes" Y if the joint past information of X and Y can better predict the future of Y than the past of Y alone does.

We further apply functional canonical correlation analysis (FCCA) to understand the covarying behavior of articulatory and acoustic measures. FCCA provides a canonical weight function for each set of time series data such that the projections of the original functional data on its weight functions are maximumally correlated.

Analysis Results

The log magnitude of G causality between articulatory and acoustic measures is shown in Table 1 with statistical significance ($p < 0.05$), where a slash means no significant G causality. We can observe that both function and content words have the same types of G causality from articulatory to prosodic characteristics in $\mathbb{C}_1$. In $\mathbb{C}_2$ and $\mathbb{C}_3$, function words have no relation between GFE and F_0, suggesting a tighter coupling between articulatory and acoustic measures for content words. In $\mathbb{C}_3$ there's an interaction between GFE and E for both content and function words, i.e., GFE "G causes" E and vice versa. We examine the difference of influence (DOI) showing that all confidence intervals of DOI at 95% significance level contain zero indicating that there's no significant difference between the influence of the two directions.

Table 1: G-Causality for Content (Con) and Function (Fun) Words in Each Category for the male speaker.

	$\mathbb{C}_1$		$\mathbb{C}_2$		$\mathbb{C}_3$	
	Content	Function	Content	Function	Content	Function
$GFE \rightarrow E$	0.003	0.0038	0.0005	0.0069	0.0019	0.0084
$E \rightarrow GFE$	/	/	/	/	0.0023	0.0089
$GFE \rightarrow F_0$	0.0049	0.0022	0.0022	/	0.0022	/
$F_0 \rightarrow GFE$	/	/	/	/	/	/

FIG. 1 demonstrates the canonical weight functions associated with the correlation coefficients in $\mathbb{C}_1$-$\mathbb{C}_3$ for content and function words. We can observe that in all cases the leading correlation coefficient of content words is higher than that of function words. We can also observe that the weight functions of articulatory and prosodic characteristics for content words vary in a similary manner, but we do not observe such covarying behavior of the weight functions for function words in most cases. These observations are consistent with those of G causality and provide support for the hypothesis that prominence realized during content words may result from a close coupling between articulatory and acoustic characteristics such as articulatory speed and F0.

ACKNOWLEDGMENTS

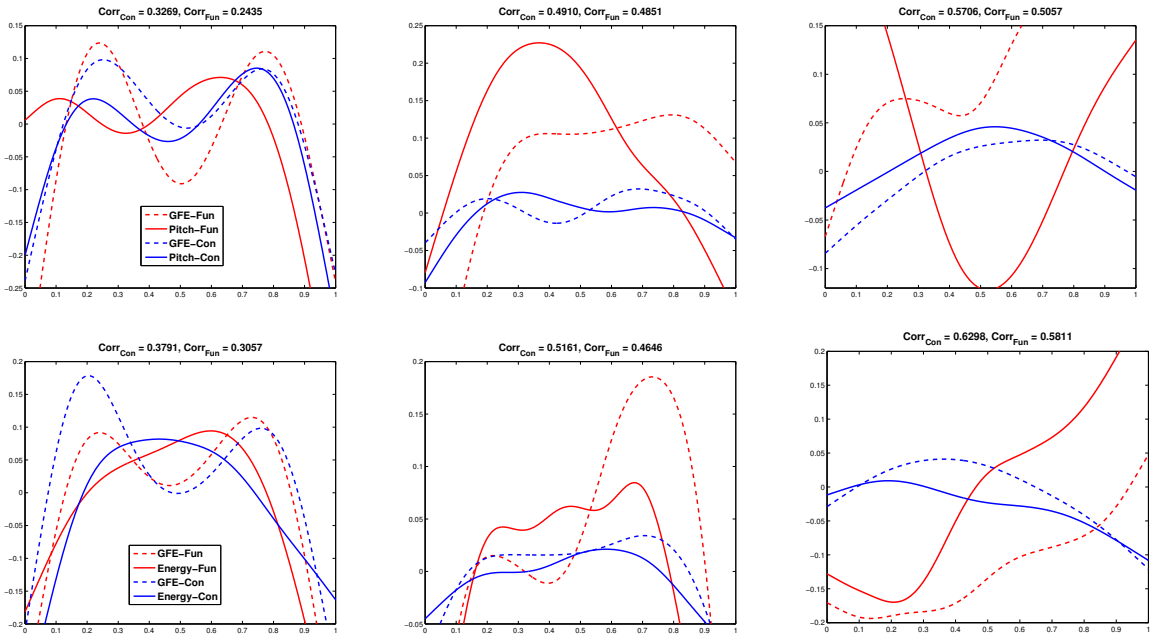
This work was supported by NIH grants DC007124 and DC03172, the LAC-USC hospital, and the USC Center for High Performance Computing and Communications (HPCC).

References

[1] T. Cho and P. Keating, "Articulatory and acoustic studies on domain-initial strengthening in korean", Journal of Phonetics **29** (2001).

[2] J. Beskow, B. Granström, and D. House, "Visual correlates to prominence in several expressive modes", in *ICSLP* (2006).

FIGURE 1: The Leading Pair of Weight Functions of articulatory and acoustic characteristics for Content and Function Words in Three Categories $\mathbb{C}_1$-$\mathbb{C}_3$ (Column 1-3) for the male speaker.



[3] D. Byrd, A. Kaun, S. Narayanan, and E. Saltzman, "Phrasal signatures in articulation", Cambridge University Press 70–87 (2000).

[4] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production", The Journal of the Acoustical Society of America **115** (2004).

[5] S. Narayanan, E. Bresch, P. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu, "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research", in *Interspeech* (Florence, Italy) (2011).

[6] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans", The Journal of the Acoustical Society of America **120**, 1791–1794 (2006).

[7] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images", Medical Imaging, IEEE Transactions on **28**, 323–338 (2009).

[8] A. Katsamanis, M. Black, P. Georgiou, and S. Goldstein, L. andNarayanan, "SailAlign: Robust long speech-text alignment", in *Workshop on New Tools and Methods for VLSPR* (2011).

[9] D. Klein and C. Manning, "Accurate unlexicalized parsing", in *Proc of the 41st Annual Meeting on Association for Computational Linguistics*, 423–430 (2003).

[10] V. Ramanarayanan, E. Bresch, D. Byrd, L. Goldstein, and S. Narayanan, "Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation", ASA (2009).

[11] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods", Econometrica: Journal of the Econometric Society 424–438 (1969).