The Effect of Word Frequency and Lexical Class on Articulatory-Acoustic Coupling

Zhaojun Yang\textsuperscript{1}, Vikram Ramanarayanan\textsuperscript{1}, Dani Byrd\textsuperscript{2} and Shrikanth S. Narayanan\textsuperscript{1,2}

\textsuperscript{1}Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA
\textsuperscript{2}Department of Linguistics, University of Southern California, Los Angeles, CA

\{zhaojuny, vramanar, dbyrd\}@usc.edu, shri@sipl.usc.edu

Abstract

Word frequency and lexical class distinction between function and content words have been shown to significantly influence word production. In this paper, we use real-time magnetic resonance imaging to investigate the effect of word frequency and lexical class on articulatory characteristics (the articulator speed) as well as acoustic characteristics ($F_0$ and short-term energy) in word production. Multiple regression analyses showed that word frequency exhibits significantly higher correlation with articulatory and acoustic factors for content words compared to function words. A Granger causality analysis uncovered a causal relationship from articulatory speed to $F_0$/energy for low-frequency content words. We further observed, using functional canonical correlation analysis, a tight coupling of articulatory and acoustic characteristics for low-frequency content words. These results support the view that word frequency distinctly influences the production of function and content words as manifested in their articulation and acoustics, as well as the dynamic coupling of these temporal streams.

Index Terms: word frequency, lexical class, articulatory-acoustic coupling, real-time MRI, speech production.

1. Introduction

An understanding of the various factors influencing word production and the variability therein is essential for developing models of speech production and recognition. In particular (and pertinent to the present paper), duration, word frequency and the lexical class distinction between the closed class of function words (prepositions, articles, pronouns, etc.) and the open class of content words (adjectives, adverbs, verbs, nouns, etc.) play a crucial role in the models of speech production [1].

The effects of word frequency on the production of words have been long studied by researchers. Early studies have pointed out the reduced articulation of frequent words [2, 3]. More recent studies have shown that less frequent words are lengthened, strengthened and have fuller forms, while frequent words are shorter in duration, and contain reduced vowels or deleted codas [1, 4, 5]. Lexical class also has been shown to influence word production. For instance, content words carry semantic meaning, and so they are generally prosodically richer than function words which carry syntactic information [6]. Research from the stuttering literature suggests that speakers are more likely to make repetition and hesitation errors on initial function words than on content words (as opposed to function words that occur after content words), and has proposed that function word dysfluencies serve to complete the plan of the upcoming content word [7]. Further support for the hypothesis that function and content words might differ in production arises from the observation that they are involved in different types of speech errors. For example, word exchanges are common in content words, while segmental shifts mainly occur on function words [8].

The distinction in word frequency effects on the production of function and content words has also been studied. Segalowitz et al. found a strong frequency effect for content words but not for function words when examining word-naming in sentence contexts [9]. Bell et al. examined the effects of word frequency and predictability on durations for function and content words, and found that frequent content words have shorter durations, while function words don’t exhibit any such characteristics [1]. They suggest that there must exist a production mechanism that associates a word’s lexical activation with its strength and rate of articulation, and that this association is implemented by a short-term coordination that moderates the pace of articulation when the progress of phonological encoding is slowed. A general motivation for a mechanism like this comes from the need for the production system to maintain temporal coordination between the conceptual/lexical and articulatory temporal streams of speech [1]. In this paper, our goal is to investigate whether such a temporal coordination exists, by studying the effect of word frequency and lexical class on articulatory as well as acoustic characteristics during word production.

Researchers have found that during speech production, people tend to make more effort to produce unfamiliar words, while they access and produce familiar words much faster [10]. Hence, we hypothesize that low-frequency content words, which are more unfamiliar than high-frequency words, might involve a tighter dynamic coupling between the articulatory and acoustic temporal streams. To test this hypothesis, we pose two questions: (1) How is word frequency related to the articulatory or acoustic signals during speech production, and further, how is this relation different for function and content words? (2) How does word frequency influence the dynamic interplay (e.g., causal interaction) of articulatory and acoustic signals for function and content words, and further, can we quantify these temporal dynamics? We address the first question using a multiple regression analysis, with word frequency as the target variable, and either articulatory or acoustic factors as predictors. For the second, we use a Granger causality analysis to uncover causal links between the articulatory and acoustic timeseries data, and functional canonical correlation analysis (FCCA) to quantify the dynamic correlation between these streams.

We use real-time magnetic resonance imaging (rtMRI) to capture the movements of the vocal tract along with synchronized audio during speech production [11]. This modality is
ideal to investigate the articulatory and acoustic dynamics and their coupling. Although there are other invasive modalities including x-ray microbeam, electropalatography, electromagnetic articulography and ultrasound that allow us to capture speech articulation at a relatively higher temporal resolution, none of these modalities offer as complete a view of all vocal tract articulators at a sufficiently high spatial resolution [12].

2. Data

We analyzed the read speech of 2 native speakers (one male and one female) of American English who spoke 460 sentences (about 3000 words) while being imaged in an MRI scanner [13]. Spoken responses and MRI videos of vocal tract articulation were recorded and time-synchronized with the audio [14]. We reconstructed the data using a sliding window technique at a rate of approximately 23.3 frames per second. Since MRI scanners generate noise, the recorded audio was post-processed using a custom noise-cancellation algorithm before analysis [15]. Further details and sample MRI movies can be found at http://sail.usc.edu/span. We automatically extracted the air-tissue boundary of the articulatory structures in each frame of video using an algorithm that hierarchically optimizes the observed image data fit to an anatomically informed object model through a gradient descent procedure [16].

3. Method

3.1. Frames of Interest

We obtained phone- and word-level alignments of the audio portion of the corpus using SailAlign, a HTK-based phonetic alignment tool [17]. We categorized the words by lexical class into function and content words in the following manner. We first parsed all words in the corpus using the Stanford Parser [18] to obtain part-of-speech (POS) tags. We then manually corrected mislabeled tags according to context. After that, we categorized each POS tagged word as a function or content word. The function word class includes articles, prepositions and conjunctions, while the content word category contains nouns, verbs, adverbs and adjectives. Note that almost all function words in our corpus are monosyllabic, so we restrict our analysis to only monosyllabic function and content words.

3.2. Articulatory and Acoustic Features

We extracted pitch ($F_0$) and short-term energy ($E$) as features of acoustic-prosodic information content. We extracted these features every 10ms with window length 30ms at the word level. We further applied a median filter to remove any spurious spikes in the pitch and energy estimates. The unvoiced and silence portions are excluded from our analysis.

As a characteristic feature of dynamic articulation, we used the ‘gradient frame energy’ (GFE) measure, which essentially captures the average speed of vocal tract articulators. GFE has been previously used for pause behavior analysis in [19]. We computed this measure as follows. Based on the extracted contour outlines of the vocal tract (see Section 2), we created binary mask images, with all pixels enclosed by these contour outlines assigned with a normalized value of 1, and the rest with 0, such that the midsagittal section of the vocal tract appears white on a black background. We then calculated the GFE measure of an MR image frame by subtracting the previous frame of binary mask image from the current one, and computing the sum of absolute values of all pixels of the resulting image. Finally, we

2. Data

We analyzed the read speech of 2 native speakers (one male and one female) of American English who spoke 460 sentences (about 3000 words) while being imaged in an MRI scanner [13]. Spoken responses and MRI videos of vocal tract articulation were recorded and time-synchronized with the audio [14].

We reconstructed the data using a sliding window technique at a rate of approximately 23.3 frames per second. Since MRI scanners generate noise, the recorded audio was post-processed using a custom noise-cancellation algorithm before analysis [15]. Further details and sample MRI movies can be found at http://sail.usc.edu/span. We automatically extracted the air-tissue boundary of the articulatory structures in each frame of video using an algorithm that hierarchically optimizes the observed image data fit to an anatomically informed object model through a gradient descent procedure [16].

3. Method

3.1. Frames of Interest

We obtained phone- and word-level alignments of the audio portion of the corpus using SailAlign, a HTK-based phonetic alignment tool [17]. We categorized the words by lexical class into function and content words in the following manner. We first parsed all words in the corpus using the Stanford Parser [18] to obtain part-of-speech (POS) tags. We then manually corrected mislabeled tags according to context. After that, we categorized each POS tagged word as a function or content word. The function word class includes articles, prepositions and conjunctions, while the content word category contains nouns, verbs, adverbs and adjectives. Note that almost all function words in our corpus are monosyllabic, so we restrict our analysis to only monosyllabic function and content words.

3.2. Articulatory and Acoustic Features

We extracted pitch ($F_0$) and short-term energy ($E$) as features of acoustic-prosodic information content. We extracted these features every 10ms with window length 30ms at the word level. We further applied a median filter to remove any spurious spikes in the pitch and energy estimates. The unvoiced and silence portions are excluded from our analysis.

As a characteristic feature of dynamic articulation, we used the ‘gradient frame energy’ (GFE) measure, which essentially captures the average speed of vocal tract articulators. GFE has been previously used for pause behavior analysis in [19]. We computed this measure as follows. Based on the extracted contour outlines of the vocal tract (see Section 2), we created binary mask images, with all pixels enclosed by these contour outlines assigned with a normalized value of 1, and the rest with 0, such that the midsagittal section of the vocal tract appears white on a black background. We then calculated the GFE measure of an MR image frame by subtracting the previous frame of binary mask image from the current one, and computing the sum of absolute values of all pixels of the resulting image. Finally, we

4. Relations between Word Frequency and Articulatory/Acoustic Signals

In this section, we examine how word frequency is related to $F_0$, short-term energy, $E$ and GFE using multiple linear regression. We extract 8 statistical functionals each of $F_0$, $E$ and GFE (see Section 3.2) for each word: mean, median, standard deviation, max, min, range, kurtosis and skewness. We conduct regression analyses with word frequency as the target, and either articulatory or acoustic factors (statistical functionals) as predictors in both function and content word cases. Table 1 presents (only) the significant articulatory and acoustic predictors of word frequency for function and content

\footnote{http://www.wordfrequency.info}

\footnote{In order to better represent the full dynamic range of word frequency values, we use a logarithmic transform of word frequency which ranges over several orders of magnitude.}
words as obtained from the regression analyses ($\alpha = 0.95$). Besides the max of GFE, the common significant factor for both function and content words, the range, standard deviation and kurtosis of GFE are all significant predictors of word frequency for content words. Furthermore, in the regression model with articulatory factors as predictors, $R^2$ is 0.153 for content words and 0.025 for function words. We also observe similar results by using acoustic functionals as predictors of word frequency, in which case $R^2$ is 0.108 for content words and 0.038 for function words. The low $R^2$ for the regression models could be due to the small number of variables (GFE, E and $F_0$) we examine. Nevertheless, the closer relation of word frequency to the articulatory or acoustic factors for content words indicates that the realization of function and content words is distinctly influenced by word frequency. We also notice that the range of $F_0$ is a significant factor of word frequency for both function and content words, which is in agreement with the finding that frequent words usually have reduced pitch range [4].

Table 1: Significant articulatory or acoustic factors regarding to word frequency for function and content words.

<table>
<thead>
<tr>
<th>Content words</th>
<th>Function words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Articulation</strong></td>
<td><strong>Articulation</strong></td>
</tr>
<tr>
<td>GFE_range</td>
<td>GFE_std</td>
</tr>
<tr>
<td><strong>Acoustic</strong></td>
<td><strong>Acoustic</strong></td>
</tr>
<tr>
<td>$E_{skewness}$</td>
<td>$F_0_range$</td>
</tr>
</tbody>
</table>

5. Articulatory-Acoustic Coupling and its Relation to Word Frequency

In the previous section, we have examined how word frequency is related to specific characteristics of articulation or the acoustic of that word. In this section, we investigate how word frequency influences the dynamic interplay between articulatory and acoustic measures for function and content words, both qualitatively and quantitatively. Based on Figure 1(a), we first categorized words into low-frequency ([-8, -4.1]) for content words and [-4, -1.8] for function words and high-frequency ([4.1, 0] for content words and [-1.8, 0]) for function words) categories. We also summarize, in Figure 1(b), information regarding the number of samples and mean duration of words in each category. Note that the classes are fairly balanced with respect to the number of samples. We also observe that low-frequency content words are significantly shorter than high-frequency ones ($p < 0.01$), which is in strong agreement with earlier studies of word duration, frequency and lexical class [1].

5.1. Causality Analysis of Articulatory and Acoustic Measures

We used Granger causality (G causality) to test the direction and magnitude of causality influence between our articulatory and acoustic measures [22]. G causality is a statistical technique for analyzing causality of time series based on linear prediction [23]. Given two signals $X$ and $Y$, consider the following two linear regression models:

$$y_1 = \sum_{i=-p}^{p} \alpha_i y_{1-i} + \epsilon(t)$$
$$y_1 = \sum_{i=-p}^{p} \beta y_{1-i} + \sum_{i=-p}^{p} \delta y_{2-i} + \epsilon' (t),$$

where $p$ is the maximum number of lags for $X$ and $Y$ in the models which is usually smaller than the number of samples. In the models above, $\alpha$, $\beta$, and $\delta$ are the regression coefficients, $\epsilon$ is the residual of predicting $Y$ from its past information, and $\epsilon'$ is the residual of predicting $Y$ based on the joint past information of $X$ and $Y$. The G causality measure is defined by taking the natural logarithm of the ratio of total variance of the residuals in the two models: $F_{X \rightarrow Y} = \ln \left( \frac{|cov(\epsilon|cov(\epsilon')|)}{|cov(\epsilon')|} \right)$, which is presented in Table 2. If the magnitude of $F_{X \rightarrow Y}$ is significantly greater than zero, which can be tested through an $F$-test, then $X$ is said to "Granger-cause" $Y$.

In each frequency category of function or content words, the feature time series, e.g., GFE, for each word is viewed as one realization. Therefore, we estimated a single model between GFE and $F_0/E$ for each category directly from the multi-realization data, instead of estimating models for each trial. Since each trial of data is required to have the same length, we uniformly sampled 50 data points for each trial from the corresponding functional data (see Section 3.3).

We did not observe any significant causal relations in either direction between articulatory and acoustic signals for both low- and high-frequency function words. Table 2 presents the significant causality influences for content words in different frequency categories ($p < 0.05$), where a slash indicates that no significant causality exists. In the low-frequency category, the articulatory speed has a significant influence on $F_0$, but we observe no such effect in the opposite direction. We also observe an interaction between GFE and speech energy. In the high-frequency category, we observe no causal relations between articulatory and acoustic measures. When we analyzed both low- and high-frequency content words together, the causal relations remain similar to those in the low-frequency category, but the magnitude of the effect is smaller. Note that the observation of causality from GFE to $F_0$ could be due to the source-tract coupling, e.g., laryngeal raising and pharyngeal expansion. Another reason might be the articulatory structures we delineated also include the glottis, although the parameter GFE in our work mainly represents the average speed of vocal tract articulators.

Table 2: Magnitudes of G-Causality between articulatory and acoustic measures for content words.

<table>
<thead>
<tr>
<th></th>
<th>Low-frequency</th>
<th>High-frequency</th>
<th>All-frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFE $\rightarrow$ $F_0$</td>
<td>0.5613</td>
<td>/</td>
<td>0.5561</td>
</tr>
<tr>
<td>$F_0$ $\rightarrow$ GFE</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>GFE $\rightarrow$ $E$</td>
<td>0.4505</td>
<td>/</td>
<td>0.4484</td>
</tr>
<tr>
<td>$E$ $\rightarrow$ GFE</td>
<td>0.4814</td>
<td>/</td>
<td>0.4740</td>
</tr>
</tbody>
</table>

5.2. Quantifying Articulatory-Acoustic Coupling through FCCA

Section 5.1 provides a qualitative analysis of interrelation between articulatory and acoustic measures. In this section, we use a function canonical correlation analysis (FCCA) based on the functional data (as described in Section 3.3) to quantify the dynamical correlation of the two types of characteristics with respect to word frequency for function and content words.

FCCA is a tool for exploring the association between a pair of functional data [20], here articulatory and acoustic streams. Given $N$ pairs of curves $(X_i,Y_i)$ defined over a finite interval $T$, the objective of FCCA is to find a pair of functions $\xi_i$ and $\eta_i$ such that the canonical variates $\int \xi_iX_i$ and $\int \eta_iY_i$ are maximally correlated, i.e.,

$$C_1 = \max_{\xi_i,\eta_i} \frac{\sum_{t \in T} (\xi'_it \cdot \eta'_it)}{\sqrt{\sum_{t \in T} \xi'_it \cdot \xi'_it} \sqrt{\sum_{t \in T} \eta'_it \cdot \eta'_it}}.$$

It is a smoothing parameter and $D^2$ denotes the second derivative. The functions $\xi_i$ and $\eta_i$ are called leading canonical variate weight functions, and $C_1$ is the leading canonical correlation. We can further look for subsidiary pairs of functions $\xi_i$ and $\eta_i$, such that the same pair of canonical variates $\int \xi_iX_i$ and $\int \eta_iY_i$ are correlated well but uncorrelated with other pairs of canonical variates. In this work, we will use the leading weight functions. The patterns of weight functions associated with canoni-
The leading canonical variate weight functions of GFE and $F0/\text{Energy}$ for content words.

Figure 2: The leading canonical variate weight functions of GFE and $F0/\text{Energy}$ for content words.

We conducted FCCA in each frequency category of function and content words respectively. We used 5-fold cross validation to select the smoothing parameter $\mu$ in each category, such that the prediction error is minimized [24]. Figure 2 shows the dynamic behaviors of the pairs of leading canonical variate weight functions for content words, where the x-axis represents time and y-axis represents magnitude of functions. Firstly, we can observe that the pairs of articulatory and acoustic weight functions for low-frequency content words, e.g., GFE and $F0$ in the left top figure, are much more tightly coupled (i.e., varying in a relative similar manner), while the weight function pairs for high-frequency content words, e.g., GFE and $E$ in the bottom right figure, do not exhibit as discernible an effect. This observation shows a tighter articulatory-acoustic coupling for low-frequency content words and adds further support to the hypothesis that people may need to make more effort to produce relatively unfamiliar words as described in Section 1. Secondly, in the low-frequency content words category, the peaks and valleys of the GFE weights are ahead of the corresponding peaks and valleys of the $F0/E$ weights (i.e., the articulatory stream leads the acoustic stream), which is not generally observed for high-frequency content words. This phase lag relationship is in agreement with the causality results presented in Section 5.1. It also suggests that this phase coupling observed in low-frequency content words might help maintain the coordination of the lexical access, articulation, and phonological encoding streams [1]. However, both of these phenomena are generally not discernibly observed for function words, as shown in Figure 3.

![Figure 3: The leading canonical variate weight functions of GFE and $F0/\text{Energy}$ for function words.](image)

There remain many exciting avenues for future work. For instance, some researchers have suggested that function words, which involve differential access as compared to content words, belong to syntactic fragments and are accessed via a feature-lookup procedure [25]. Our analysis using articulatory data allows us to find support for, and consequently extend this model. Note that some potential confounds which we did not control for include phrase position and prosodic conditions. Additionally, we only looked at a small subject pool of 2 subjects. These factors will be included to extend the analysis for future research.

6. Conclusion and Discussion

This paper investigates the relationship between the word frequency and the realization of function and content words in terms of articulatory and acoustic measures. We hypothesize that low-frequency content words, that are more unfamiliar than high-frequency words, might involve a tighter coupling between the articulatory and acoustic temporal streams. To test this hypothesis, our work has attempted to answer two specific questions: (1) how is word frequency related to the articulatory speed or acoustic measures ($F0$ and short-term energy) respectively? (2) how does the word frequency affect the dynamic interplay of articulatory and acoustic signals in word production? Multiple regression analyses show that word frequency is more correlated with articulatory and acoustic signals for content words, supporting the view that word-familiarity affects the realization of words and such an influence differs for function and content words. Furthermore, the causal relationship observed between GFE and $F0/E$ Energy for low-frequency content words adds support to the hypothesis that the word planning in these cases involves a close coordination between lexical access, articulatory planning, and phonological encoding processes [1]. In addition, we observe a tighter coupling of articulatory and acoustic weight functions and a stronger articulatory-acoustic correlation for low-frequency content words, lending further support to our hypothesis that the word familiarity affects the realization of function and content words not only in terms of articulation or acoustics individually, but also in a dynamic sense via articulatory-acoustic coupling.

There remain many exciting avenues for future work. For instance, some researchers have suggested that function words, which involve differential access as compared to content words, belong to syntactic fragments and are accessed via a feature-lookup procedure [25]. Our analysis using articulatory data allows us to find support for, and consequently extend this model. Note that some potential confounds which we did not control for include phrase position and prosodic conditions. Additionally, we only looked at a small subject pool of 2 subjects. These factors will be included to extend the analysis for future research.

Table 3: The leading canonical correlation $C_1$ between articulatory and acoustic measures for function and content words.

<table>
<thead>
<tr>
<th>Content words</th>
<th>Function words</th>
<th>Low-freq</th>
<th>High-freq</th>
<th>Low-freq</th>
<th>High-freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GFE</td>
<td>0.3118</td>
<td>0.2392</td>
<td>0.2016</td>
<td>0.2305</td>
<td></td>
</tr>
<tr>
<td>F0-GFE</td>
<td>0.3216</td>
<td>0.3007</td>
<td>0.2279</td>
<td>0.2235</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 presents the corresponding leading canonical correlations $C_1$ between the articulatory speed GFE and acoustic measures $F0$ and $E$. The canonical correlations for content words in low-frequency category are higher as compared to high-frequency content words. High-frequency function words have a higher correlation between GFE and $E$ as compared to low-frequency function words. We generally observe that functional correlations for content words are higher than those for function words. These results are consistent with the observation of weight functions obtained above.

3. Further analysis

In addition to the leading canonical variate weight functions, we conducted multiple regression analyses to investigate the relationship between word frequency and articulatory-acoustic measures. We hypothesized that word frequency affects the articulatory and acoustic characteristics of words, and that this relationship is stronger for content words. To test this hypothesis, we performed multiple regression analyses on the leading canonical variate weight functions. The results show a significant correlation between word frequency and the articulatory-acoustic measures, with content words exhibiting a stronger correlation than function words. This finding supports the view that word familiarity affects the realization of words.

4. Discussion

The results of this study suggest that word frequency plays a crucial role in the realization of words, with content words being more affected than function words. These findings are consistent with previous research on word frequency and its impact on articulatory-acoustic characteristics [26]. The stronger correlation between word frequency and articulatory-acoustic measures for content words is likely due to the fact that content words are more unfamiliar and may require more effort to be produced correctly. This finding has important implications for speech production models, as it suggests that accounting for word frequency is necessary to accurately model the realization of words.

5. Conclusion

In conclusion, this study provides evidence for the relationship between word frequency and articulatory-acoustic measures, with content words being more affected than function words. The findings suggest that word familiarity affects the realization of words, with content words being more affected than function words. These results have implications for speech production models and highlight the importance of accounting for word frequency in these models. Future research should continue to investigate the role of word frequency in speech production, as well as explore the potential confounds that may affect this relationship.
7. References


