LANGUAGE REDUNDANCY EFFECTS ON F0: A PRELIMINARY CONTROLLED STUDY

Cong Zhang[1], Catherine Lai[2], Ricardo Napoleão de Souza[2], Alice Turk[2], Tina Bögel[3]

[1]Newcastle University, UK [2]University of Edinburgh, UK [3]University of Konstanz, Germany
cong.zhang@newcastle.ac.uk; c.lai@ed.ac.uk; a.turk@ed.ac.uk;
R.N.deSouza@ed.ac.uk; Tina.Boegel@uni-konstanz.de

ABSTRACT

Previous research suggests that words with a high level of language redundancy (i.e. recognition likelihood from familiarity and predictability based on syntactic, pragmatic, and semantic factors) have reduced acoustic salience, such as shorter duration and reduced vowels. The Smooth Signal Redundancy Hypothesis proposes that acoustic salience is controlled via prosodic structure, and makes the prediction that parameters such as fundamental frequency should also be affected by language redundancy. This study investigates the relationship of F0 with lexical frequency, together with bigram (verb-adjective or adjective-noun) frequency and the ratio between these two bigram frequencies. Results from a carefully controlled experiment with quadruplets of minimal pairs suggests that language redundancy can affect fundamental frequency in English.

Keywords: Smooth Signal Redundancy Hypothesis, f0, frequency effects, prosodic structure, English

1. INTRODUCTION

In order to achieve efficient communication, speakers manipulate acoustic salience based on how redundant given linguistic units are in discourse [1]. Language redundancy encompasses a range of pragmatic, lexical, syntactic, and semantic factors. Acoustic salience refers to segment duration and spectral properties associated with hyperarticulation, as well as to acoustic properties associated with more salient prominence markers and stronger prosodic boundaries, including f0. Language redundancy correlates inversely with acoustic salience, so that the more redundant the linguistic item, the less acoustically salient it is, and vice versa [1, 2, 3, 4]. For instance, a speaker might highlight a rare or unexpected word (i.e., less redundant) by making it longer and/or by making it stand out in the speech stream through more pronounced boundary markings [5]. Since acoustic salience varies inversely with language redundancy, the information conveyed in an utterance is distributed more evenly, thereby maximising its likelihood of recognition. The Smooth Signal Redundancy Hypothesis [1, 5]proposes that prosodic structure is used to control acoustic salience.

While previous research suggests robust effects of language redundancy on duration measures [1, 6, 7], much less is known about its local effects on f0 (but see [4]). The current study tests the extent to which speakers manipulate f0 on strings of words that differ in their language redundancy profile. Given that f0 movement is one of the acoustic cues English speakers use to cue prosodic prominence via intonation contours (e.g., [8]), the Smooth Signal Redundancy Hypothesis (SSRH) predicts that redundancy should affect f0 in systematic ways.

In a study of spontaneous American English, [4] found that contextual plausibility (i.e., a measure of redundancy) affected f0 values as predicted by the SSRH: lower redundancy yielded overall higher f0 values. On the other hand, discourse mention and focus status showed less clear results, suggesting that redundancy might affect f0 differently from duration. Unlike the current study, however, lexical frequency measures were not manipulated in that investigation.

Using data obtained under tightly controlled experimental conditions, this study investigates the relationship between acoustic salience measures (prosodic prominence, boundary tone) and three measures of language redundancy: lexical frequency, bigram frequency (verb-adjective, adjective-noun), and the ratio between those two bigram frequencies.

2. METHODS

2.1. Materials

The materials used in this study were originally designed to examine the relationship between duration and language redundancy in [7]. Recordings were made in a sound-treated studio at the University of Edinburgh, at a sampling rate of 44.1 kHz, 16-bit.

A total of 23 participants (14F) participated in the reading task. All participants were students at the University of Edinburgh and received compensation.

Fourteen sets of well-balanced quadruplets were created for the study such as those exemplified in (1). Each utterance in a set included a Verb-Adjective-Noun (V-A-N) sequence. Each participant was randomly presented with the target sentences twice.

Three frequency measures were used as indicators for (non-acoustic) language redundancy:

a. Lexical frequency: The verbs and the nouns selected had either frequent (f) or infrequent (i) usages. The lexical frequencies were obtained from the WebCelex's Cobuild corpus [9]. For verbs, f corresponds to a raw number of occurrences above 2000, whereas i indicates fewer than 200. For nouns, f corresponds to raw frequencies over 3000, whereas i indicates raw frequencies below 100. A substantial buffer zone (V: 200-2000; N: 100-3000) was deliberately selected to ensure that the two categories were distant enough in their lexical frequencies. Frequent and infrequent verbs in each set both shared an identical rhyme, e.g. *make* vs *rake* in (1). On the other hand, frequent and infrequent nouns shared the same onsets and nuclei, e.g. *fields* vs *fiefs* in (1). The adjective between the verb and the noun was the same for all four in a quadruplet. In the examples below, frequency codes are labelled as subscripts on the verbs and the nouns.

(1) A quadruplet with frequent(f) and infrequent(i) verbs and nouns

ff: Whatever you **make**$_f$ **clean fields**$_f$ should be a priority
fi: Whatever you **make**$_f$ **clean fiefs**$_i$ should be a priority
if: Whatever you **rake**$_i$ **clean fields**$_f$ should be a priority
ii: Whatever you **rake**$_i$ **clean fiefs**$_i$ should be a priority

All sentences in the quadruplets were syntactically ambiguous. As a result, the adjective could be prosodically grouped with either the preceding verb (VA%N) or with the following noun (V%AN). No punctuation was given in the experiment to indicate prosodic phrasing to allow for participants' unbiased boundary placement. In the ICE-GB corpus [9] and the Brown corpus [7], the V%AN structure is much more frequent (ca. 77%) than the VA%N counterpart (ca. 4%).

b. Bigram frequency: Word bigram frequency was obtained through Google for the VA sequences and the AN sequences respectively. The bigram frequencies were categorized as "`high`" for frequencies above 60% of the total range (9 to 45 million occurrences), and "`low`" for data below 40%. A 20% buffer zone was kept to avoid borderline frequency categories.

c. Ratio of bigram frequencies: The ratio of the two bigram frequencies was calculated by dividing the bigram frequency of VA by the bigram frequency of AN, i.e. RATIO-BI = $Freq_{(VA)}$ /$Freq_{(AN)}$. The methods for normalizing and parting data followed the bigram frequencies'. The frequency ratio was coded with the higher value between $Freq_{(VA)}$ and $Freq_{(AN)}$.

2.2. Analysis

Participants were free to choose the parsing that they preferred for each utterance, for instance by placing the prosodic boundary either between V and AN (V%AN) or between the VA and N (VA%N). We only analyzed the utterances where all four utterances and both repetitions in a set were produced with the same prosodic phrasing. A total of 344 utterances (produced by 11 speakers, from 11 quadruplets) were then selected for the final dataset.

2.2.1. Data annotation

Segments in all utterances were automatically labelled using a customised version of the Montreal Forced Aligner [10]. Sonorant segments in target words (V, A, N) of each utterance were labelled on a different tier and used for data extraction. The intonation of all utterances was labelled following the original Tone and Break Indices [11] to inspect pitch accents, phrase-tones and boundary tones used in the production.

2.2.2. F0 extraction

Three analyses were conducted to examine language redundancy effects on f0. All f0 values were extracted using Praat [12] with a range of 75 Hz to 400 Hz using "Get mean" function in Praat.

Sonorant intervals of each target word (e.g. /meɪ/ in "make", /liːn/ in "clean") were divided into three equal portions for the analyses (henceforth *initial third*, *second third*, *final third*). Our test words were monosyllabic, and when pre-boundary, any pitch accents, phrase-tone, and boundary tone would all occur on the same syllable. Our rationale for dividing the syllable into three was that the f0 of the initial third might correspond more closely to pitch-accent-related f0, and the f0 of the last third might correspond more closely to boundary-related f0. We acknowledge that dividing a syllable into three parts based on f0 is a simplification and does not capture the full complexity of intonation patterns. However, we are using this division to provide some preliminary insights.

2.2.3. *Statistical analysis*

The data were divided into two datasets according to their prosodic grouping, i.e. V%AN (296 utterances in total) and VA%N (48 utterances in total). The data were analyzed using linear mixed effect models [13] in R [14]. The base model only included the predicted factors and their interactions, i.e. the lexical frequencies of V and N and their interaction (V-FREQ*N-FREQ). The interaction term was removed when insignificant and if a reduced model was a better fit (i.e. with a lower AIC). The duration of sonorant interval (SONORANT-DUR) was also added to test whether the f0 and duration are related. A full model includes all relevant factors as the main effects, including the lexical frequencies of V and N and their interaction (V-FREQ*N-FREQ), the duration of sonorant interval (SONORANT-DUR), the bigram frequency for VA and AN respectively (VA-FREQ, AN-FREQ), as well as the ratio between VA and AN (RATIO-BI). SPEAKER and the index of QUADRUPLET were used as random intercepts; when models failed to converge, QUADRUPLET was removed. The dependent variable in the current analysis is MEAN F0. We tested the mean f0 of the initial, second, and final third of the verb and the noun respectively in different statistical models.

3. RESULTS

Tune types did not seem to be associated with word frequency. In the V%AN dataset, the V and the N mainly had a !H* pitch accent, with a falling or a flat boundary. The A mostly had a H* pitch accent. The most popular tune was V [!H* + H-L%] + A [H*] + N [!H* + H-L%], which was used in 74 utterances (30% of the dataset). This tune had a flat contour for both three target words, making the entire utterance fairly monotonous and flat. The VA%N dataset was much smaller and the most used tune was V [!H*] + A [!H* H-L%] + N [H* + H-L%]. A small percentage of the tunes end with a H% boundary or edge tone, including 28.5% in V, 6.4% in A, and 4.7% in N across both datasets. These instances were excluded from the analysis to keep the analysis unified and the predictions consistent.

If SSRH holds for f0, we would expect a higher f0 for infrequent words when the pitch accents and boundary tones are high, and lower f0 when the tones are low. Our prediction for !H* was unclear.

The following results are based on the best-fit models. None of the full models turned out to be the best-fit model so the bigram frequency measures, including VA-FREQ, AN-FREQ and RATIO-BI, were not significant predictors for the dependent variables.

For the VA%N dataset, the statistical analysis showed that none of the frequency effects was significant, possibly due to the small number of instances in this dataset (48 instances). Therefore we will not report further about this dataset.

3.1. Mean f0 of initial third

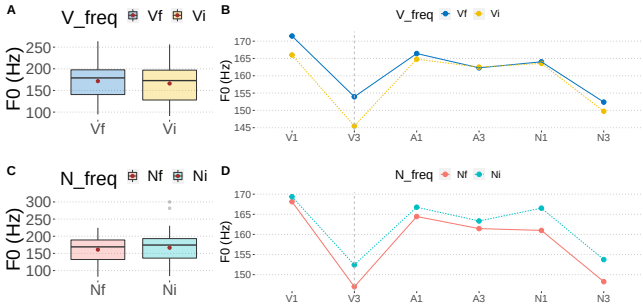

Figure 1: A & C: Mean f0 of the 1st third of V for each V-FREQ group (A) and for each N-FREQ group (C). Red dots: mean f0. **B & D**: Schematized lines by V-FREQ (B) and N-FREQ(D). V1/A1/N1: mean f0 of the initial third in V/A/N; V3/A3/N3: mean f0 of the final third in V/A/N.

In the V%AN dataset, the results showed two significant predictors on the initial third of the verb and the initial third of the noun. For the initial third of the V, the mean f0 was lower when the V was infrequent than when it was frequent (β = -7.989, SE = 3.106, t = -2.572, $p < 0.05$), as shown in Fig.1A. For the initial third of N, the mean f0 was significantly higher when the noun frequency was low ($\beta = 4.342$, SE = 2.091, t = 2.076, $p < 0.05$), as shown in Fig.1C. Fig.1B divides the data into two groups by V-FREQ, showing a difference of around 5 Hz in the verb (V1). Similarly, Fig.1D divides the data into two groups by N-FREQ, and exhibits a difference of approximately 5 Hz in the noun (N1). Both results indicated a link between the F0 value for the first third of the sonorant interval and word frequency, and thus provided certain insights for SSRH. However, because the pitch accent on V1 was downstepped, we did not have a clear hypothesis for the direction of the effect. The effect of mean f0 for N was in line with our expectations, i.e. higher f0 on a H* when N was infrequent.

3.2. Mean f0 of final third

Following the duration analysis in [7], in which the effect of V-FREQ led to a significant increase of 15 ms in the pre-boundary V coda when the verb was infrequent, we hypothesize that pre-boundary f0 is also affected by the frequency of the pre-boundary

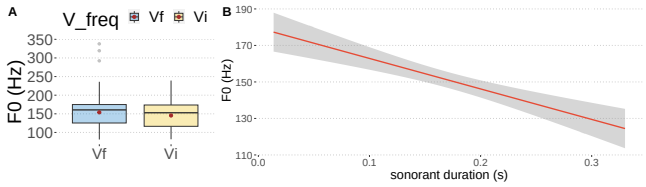

Figure 2: **A**: Mean f0 of the final third of V for each V-FREQ group. Red dots show the mean f0 difference of each condition. **B**: Regression line between the mean f0 and the duration of sonorant segment in the final third of N V-FREQ. Grey area indicates 95% confidence interval.

word.

In V%AN, the results show significant changes brought by the frequency measures and other acoustic measures on the V and the N respectively. For the final third of V, the mean f0 was significantly lower when V-FREQ was infrequent (β = -8.448, SE = 3.062, t = -2.759, $p < 0.01$), as shown in Fig. 2A. Fig.1B illustrates that when V is infrequent, the boundary tone indicated by V3 on the yellow dotted line is approximately 8 Hz lower than in the frequent V condition (V3 on the blue solid line). This is consistent with the SSRH prediction of a stronger prosodic boundary, i.e. an even lower boundary tone, after a less frequent word. For the final third of N, the mean f0 was also lower as the duration of the sonorant interval increased (β = -90.536, SE = 33.441, t = -2.707, $p < 0.01$), as in Fig. 2B. This clear inverse correlation between temporal information and f0 demonstrates further support for SSRH.

3.3. Global analysis

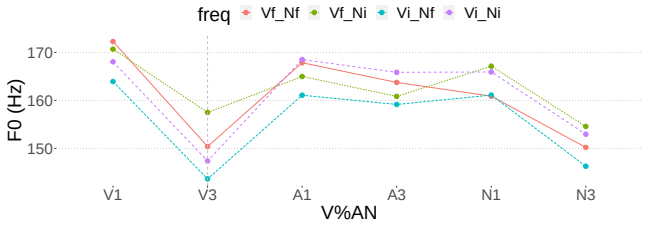

Figure 3: Schematized lines for V%AN dataset by verb-noun frequencies. V1/A1/N1: mean f0 of the initial third in V/A/N; V3/A3/N3: mean f0 of the final third in V/A/N. Grey dashed vertical lines indicate prosodic boundaries.

Fig. 3 shows schematized lines using the mean f0 values of the initial and final thirds of V, A, and N. Different lines demonstrate different combinations of verb frequencies and noun frequencies.

The results from the V%AN dataset showed a mixed picture. While both the final third of V (V3) and the initial third of N (N1) were inversely correlated with frequency measures, i.e. the lower the V-FREQ, the higher the f0 values, the mean f0 of the initial third of V (V1) changes positively with the frequency measures, i.e. the f0 was lower when the V-FREQ was low. There are several possibilities that can explain V1's pattern. The first cause for the infrequent V (Vi) being lower in the initial third (V1) than the frequent ones (Vf) may be that the low boundary tone in the final third of the V (V3) has brought down the f0 of the entire word since the words are quite short in this dataset. An additional analysis of the mean f0 of the second third of the verb showed that the f0 of V was lower for the infrequent verb by around 5 Hz than the frequent verbs (β = -5.033, SE = 2.250, t = -2.237, $p < 0.05$), consistent with the results from V1 and V3. This also suggests that the extra acoustic salience is indicated more by a much lower boundary tone than a higher pitch accent. Moreover, as mentioned previously, V having a !H* pitch accent further complicated the issue. When the pitch accent is a H*, the prediction is that the mean f0 will be higher to increase prominence[4]; however, the prediction is less clear when the pitch accent is a !H* since its F0 value may be more easily influenced by other surrounding tones. Another contributing factor may be the general prosodic composition in English. In Fig. 3B, the Vf_Nf line (red solid line) is a typical intonation contour for a (!)H* + H-L% + H* + H-L%, i.e. a high pitch accent with a falling boundary and another pitch accent with a falling boundary. In contrast, Vf_Ni seems more likely to have a narrow focus (L+H*) on the infrequent noun. Last but not least, the f0 contour on the verb was also affected by the upcoming noun: the mean f0 of the second third of the verb was also influenced by N-FREQ (β = 4.473, SE = 2.250, t = 1.988, $p < 0.05$), showing that the second third of V was lower in f0 when the frequency of the upcoming noun was higher. This highlights the need for more understanding of global phrasal patterns and long-distance coarticulation in intonation.

4. DISCUSSION AND CONCLUSIONS

In summary, this study offers useful preliminary observations regarding the relationship between frequency measures and f0 in experimental English data. In line with [4], the current results provide some tentative support for the *Smooth Signal Redundancy Hypothesis* and suggest that language redundancy does affect f0 in some ways. It also confirms that different acoustic salience measures, such as duration and f0, influence each other in a complementary manner. However, further insights might emerge with an improved design.

5. ACKNOWLEDGEMENTS

We gratefully acknowledge funding from AHRC-DFG Grant No. AH/W010801/1, to A. Turk , T. Bögel and C. Lai.

6. REFERENCES

[1] Aylett, Matthew and Turk, Alice, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, vol. 47, no. 1, pp. 31–56, 2004.

[2] Lindblom, Björn, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*. Kluwer Academic Publishers, 1990, pp. 403–439.

[3] Bell, Alan, Brenier, Jason, Gregory, Michelle, Girand, Cynthia, and Jurafsky, Dan, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.

[4] Turnbull, Rory, "The role of predictability in intonational variability," *Language and Speech*, vol. 60, no. 1, pp. 123–153, 2017.

[5] A. Turk, "Does prosodic constituency signal relative predictability? A Smooth Signal Redundancy hypothesis," *Laboratory Phonology*, vol. 1, no. 2, pp. 227–262, 2010. [Online]. Available: https://doi.org/10.1515/labphon.2010.012

[6] Aylett, Matthew and Turk, Alice, "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3048–3058., 2006.

[7] T. Bögel and A. Turk, "Frequency effects and prosodic boundary strength," in *International Congress of Phonetic Sciences: ICPhS2019*, 2019, pp. 1014–1018.

[8] Ladd, D. Robert, *Intonational Phonology*. Cambridge University Press, 2008.

[9] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "WebCelex," 2001, online resource: http://celex.mpi.nl/.

[10] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[11] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The Original ToBI System and the Evolution of the ToBI Framework," in *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 2005.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. Version 6.2.23," 2022, online resource: http://www.praat.org/.

[13] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, pp. 1–48, 2015. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v067i01

[14] R Core Team, "R: A language and environment for statistical computing," Vienna, Austria, 2022. [Online]. Available: http://www.r-project.org/