10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Revisiting Neutral Tone in Mandarin Broadcast News Speech

Chenzi Xu

University of Oxford, UK
chenzi.xu@ling-phil.ox.ac.uk

Abstract

This study examines the duration and fundamental frequency (f_0) contour of neutral tones in the 1997 Mandarin Broadcast News Speech Corpus. 7208 disyllabic occurrences with a lexical tone syllable preceding a neutral tone syllable and 180 trisyllabic occurrences in which two consecutive neutral tone syllables follow a lexical tone syllable from 21 speakers were analysed using orthogonal polynomial and linear mixed effects models. Instead of the oversimplified claim that neutral tone syllables are shorter than other syllables, the results suggest that the length of a neutral tone syllable is sensitive to its syllable structure. The results also capture a converging low pitch target with large variance for Mandarin neutral tones, having visualised the distribution of f_0 variation of neutral tones.

Index Terms: neutral tone, Mandarin, fundamental frequency contour

1. Introduction

Despite abundant studies on the four canonical lexical tones of Standard Mandarin, the phenomenon of neutral tone is often understudied. Neutral tones are not limited to a few particles that received some scholarly attention such as the nominaliser or possessive particle *de* /tə/, but occur frequently in many disyllabic words and phrases.

Neutral tone syllables such as (1) from [1], characterised by reduced duration and vowel contrasts, show how stress plays a role in Standard Mandarin [2]. In (1), when the syllable /faŋ˩/ is in the unstressed position in a disyllabic word *dì fāng*, the velar nasal /ŋ/ is pronounced without the velar closure and nasally coarticulated with the vowel, the back low unrounded /ɑ/ is realised as a mid-central vowel [ə], and has a neutral tone rather than a high tone.

(1) *fāng* /faŋ˩/→**[fə̃]** in [ˈti˥˩ **fə̃**] *dì fang* 地方 'place'

A neutral tone syllable is a weak syllable exhibiting contextually conditioned pitch realisation. It never occurs in initial positions, mostly enclitically, where its pitch patterns tend to heavily depend on the preceding syllable [3, 4]. The contextual dependency of pitch realisation of the neutral tones revealed by impressionistic observations and earlier acoustic studies based on scanty data has led to few attempts on the phonological specifications of neutral tone as an independent tonal category. This study adopts the corpus approach and examines the duration and f_0 contour of a relatively large amount of neutral tone syllables in various contexts in broadcast news speech.

2. Method

2.1. Sound Recordings

This study used the 1997 Mandarin Broadcast News Speech corpus consisting of broadcast news recordings of a single channel and 16,000 Hz sample frequency, released by the Linguistic Data Consortium. The speech data were automatically time aligned by syllables with corresponding transcripts and tone category labels, which were obtained from Jiahong Yuan [5]. Excluding recordings that involve background noise and music, non-standard accent, and multiple speakers, we examined utterances from 21 speakers, 13 male and 8 female.

Two types of utterances were examined: disyllabic occurrences [X-N], in which a neutral tone syllable represented by N follows a non-neutral tone syllable X, which can be any of the four lexical tones of Mandarin (i.e. H, LH, L, HL); and trisyllabic occurrences [X-N_1-N_2], in which two consecutive neutral tone syllables follow a non-neutral tone syllable. [X-N] occurrences containing five of most frequently-used functional morphemes including /tə/, /lə/, /mən/, /ʈʂə/ and /tsɻ/ were selected for analysis due to the fact that their corresponding lexical tone variant is hardly available in such occurrences. Extremely short clips that generate no pitch track or clips that are voiceless throughout the utterance were also excluded from the pitch analysis, but they were included in the duration analysis. The final dataset for pitch analysis comprised of 7208 tokens of neutral tone syllables for [X-N] disyllabic utterances, and the number of token after the H, LH, L and HL tone syllable is 1654, 1750, 1210, and 2594 respectively. There were 180 tokens of trisyllabic utterances, and the number of token after the H, LH, L and HL tone syllable is 49, 43, 56, and 32 respectively.

2.2. Acoustic Measures

The automatic forced aligned annotation contains temporal information of the boundaries of each syllable and thus the duration of each syllable. f_0 estimates (in Hertz) in 10 millisecond intervals of voiced regions of all the recordings of each speaker were obtained using the *get_f0* program from the ESPS[1] package developed by Entropic Research Laboratory. Based on these f_0 estimates, the average f_0 of each speaker was calculated.

Having obtained the sound clips of neutral tones in disyllabic and trisyllabic utterances, the f_0 contours of the selected neutral tone syllables were generated using the ESPS *get_f0* program. By using the *polyfit* function in GNU Octave

[1] Entropic Signal Processing System, release of the Phonetics Lab, University of Oxford. Downloaded from: http://www.phon.ox.ac.uk/releases

10.21437/SpeechProsody.2020-104

[6], an open-source version of Matlab, a quadratic polynomial was fitted to each f_0 contour of the neutral tone syllable in disyllabic utterances and [N_1-N_2] part of the trisyllabic utterances, deriving an effective model of the pitch contour shape of a neutral tone syllable and smoothing out the small bumps or pitch perturbations that are considered unimportant. A best-fit polynomial model is demonstrated in Figure 1.

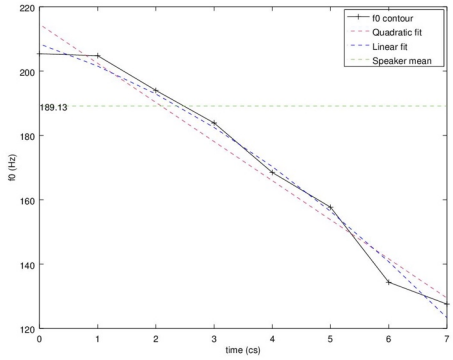

Figure 1: *f_0 contour and polynomial model of token* /tə/ *in* /ɛin˥ tə/ *'new'.*

2.3. Analytical Models

In order to compare the shape of these f_0 contours, f_0 frequencies in Hz and times in cs were then normalised. f_0 estimates of each phrase were normalised to the corresponding speaker mean $\overline{f_s}$ and centred around zero using the following formula:

$$f' = \frac{f_0}{\overline{f_s}} - 1 \quad (1)$$

The corresponding time values were linearly scaled to the range of -1 to 1, following (2)

$$t' = \frac{2(t-\frac{L}{2})}{L} \quad (2)$$

where L is length of the voiced regions of a token. The f_0 curve of a token is now recast as the function $f'(t')$ in the normalised domain. The range -1 to 1 was chosen because it enabled the use of orthogonal Legendre polynomials that are defined over the interval [-1,1] to model the curve.

Informed by the method used in [7, 8], a second-order Legendre polynomial model $f'_M(t')$ (3) was used to model each normalised curve, given that orthogonal polynomials derive coefficients with minimal correlations and their coefficients are uniformly sensitive throughout the utterance.

$$f'_M(t') = c_1L_1(t') + c_2L_2(t') + c_3L_3(t') \quad (3)$$

The three coefficients, i.e. c_1, c_2, and, c_3 obtained from each model characterise the f_0 contour in three aspects: AVERAGE HEIGHT, SLOPE, and CURVATURE. They were subjected to further Linear Mixed-Effects (LME) models analysis.

The independent fixed-effect predictor was the preceding tone category, and the random effect was speakers. The five different neutral tone particles in disyllabic utterances were also included as a random effect. Only random intercepts were included in the models because random slopes were not properly supported by the data and the models failed to converge if adding them. The analysis was performed in R [9] using the LME4 package [10] and LMERTEST package [11].

3. Results

3.1. Duration

Table 1 lists the mean and standard deviation of the duration of both syllables of the disyllabic utterance by the neutral tone suffix, and Figure 2 draws the corresponding kernel density function, a smoothed alternative to histogram for continuous data. Boxplots show the interquartile range and the median of the duration of all lexical tone and neutral tone syllables. From Table 1, the overall mean of the neutral tone is about 72% of the mean of other lexical tones. In order to examine whether the mean difference in syllable duration is significant, a paired-samples upper-tailed t-test was conducted to compare the duration of the neutral tone with its preceding lexical tone in the same utterance. There was a significant difference in the duration of lexical tones (μ = 0.158, σ = 0.046) and neutral tones (μ = 0.114, σ = 0.049); t(8416) = 68.169, p < .001.

Table 1 also suggests that the shorter duration of neutral tone may relate to the syllable structure. Among the five neutral tone syllables, suffix men /mən/, the only one with a nasal coda in the syllable structure, has the reversed result where the neutral tone syllable is on average significantly longer (t(541) = -15.471, p <.001) than its preceding lexical tone syllable, which consist of mostly (82.3%) CV syllables such as /tʰa˥/ (217/542) and /wo˩/ (229/542). Although [12] reported that hardly any significant difference between the duration of CVN syllables and CV syllables was found, we cannot yet hastily conclude that it is the neutral tone status instead of simpler syllable structure that leads to the relatively shorter duration.

Table 1: *Mean and Standard Deviation of the duration (s) of syllables in the [X-N] disyllabic utterances*

Suffix	Lexical tone	Neutral tone
-de /tə/ 的	0.161 (σ= 0.045)	0.109 (σ= 0.047)
-le /lə/了	0.157 (σ= 0.040)	0.122 (σ= 0.049)
-men /mən/们	0.109 (σ= 0.033)	0.141 (σ= 0.052)
-zhe /ʈʂə/着	0.167 (σ= 0.045)	0.129 (σ= 0.055)
-zi /tsɻ/子	0.234 (σ= 0.057)	0.185 (σ= 0.061)
Overall	**0.158** (σ= 0.046)	**0.114** (σ= 0.049)

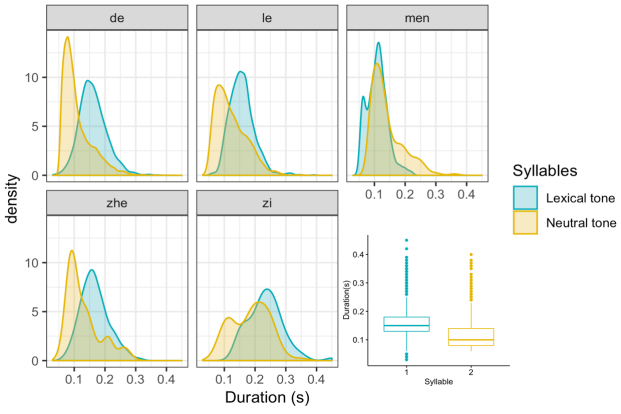

Figure 2: *Kernel density functions and boxplots for the duration (s) of syllables in [X-N] disyllabic utterances.*

Table 2: *Mean and Standard Deviation of the duration (s) of syllables in trisyllabic utterances and two paired-samples upper-tailed t-tests results*

Lexical Tone	Neutral Tone 1	Neutral Tone 2
0.160 (σ= 0.061)	0.122 (σ= 0.033)	0.108 (σ= 0.051)
t(198) = 8.584, p <.001		
	t(198) = 3.665, p <.001	

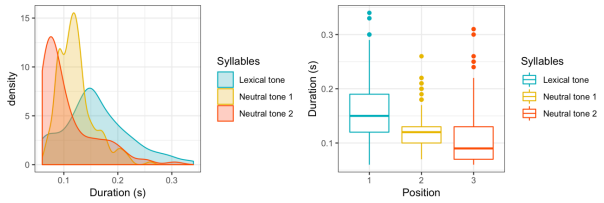

Figure 3: *Kernel density functions and boxplots for the duration (s) of syllables in the [X-N$_1$-N$_2$] trisyllabic utterances.*

Table 2 and Figure 3 present the durational information of syllables in trisyllabic utterances. From the initial syllable to the last syllable in an utterance, the average duration of each syllable decreases. The downward shifting boxplots in Figure 3 capture the same trend. In Figure 3, the graph to the left with three non-overlapping peaks shows the distribution of the duration of syllables at different positions. A paired-samples upper-tailed t-test was conducted to compare the duration of the middle neutral tone with its preceding lexical tone and the duration of the two neutral tones in the same utterance. The *p* values are less than .001, which indicate that in these utterances the last neutral tone on average is significantly shorter than the middle neutral tone and the middle neutral tone on average is significantly shorter than the preceding lexical tone.

The majority of the last neutral tone syllables (76.9%) in these utterances are *de* /tə/, and about half of the middle neutral tone (47.7%) syllables are *men* /mən/, the relatively long neutral tone from our disyllabic data. This may explain why the second neutral tone on average is shorter than the first neutral tone in the trisyllabic dataset when there are two consecutive neutral tones.

3.2. patterns of monosyllabic neutral tone particles

Table 3 summarises the significance level of aspects of shape predicted by the preceding lexical tone category in the LME models. The p values are all less than .001, which suggest the preceding tone exerts some statistically significant influence on the shape of contour of the following neutral tone.

Table 3: *Fixed effects in the linear mixed-effect models for aspects of pitch contour of monosyllabic neutral tone particles*

Shape	Predictor	df1	df2	F	*p*
HEIGHT	Tone X	3	7112.5	59.011	**<.001**
SLOPE	Tone X	3	7153.7	371.06	**<.001**
CURVATURE	Tone X	3	7195.9	402.14	**<.001**

Model: Shape ~ Preceding Tone Category + (1| Speaker) + (1|Particle).

Figure 4 illustrates the aspects of shape represented by the Least Squares means of coefficients in the LME models by reproducing the modelled contours of neutral tone. Instead of assigning a pitch level impressionistically in previous studies, it offers a acoustically-based much detailed prototype of neutral tone contour in different tonal contexts. The contours start at various pitch heights due to the different preceding tones, and end within a narrower pitch range. Figure 5 show the distribution of normalised f_0 at the beginning and the end of the voicing part of the neutral tone of all the utterances. To the right of the boxplots, kernel density functions and unimodal Gaussian distributions are shown to model the pitch variation at this time point.

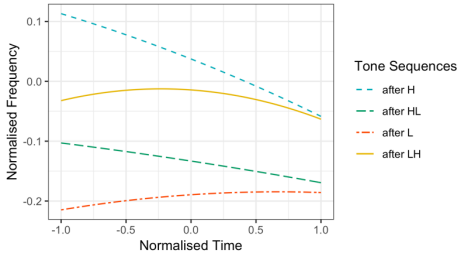

Figure 4: *Simulation of single neutral tone contours when preceded by different tones.*

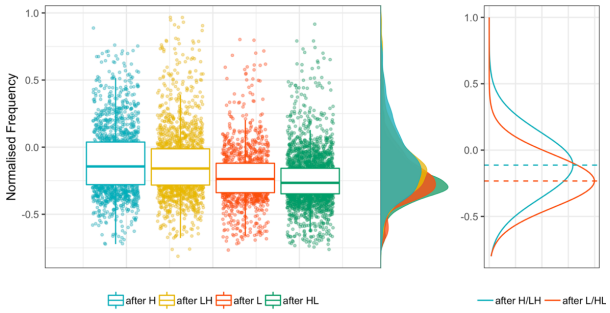

Figure 5: *Distribution of modelled ending pitch of neutral tones in [X-N] disyllabic utterances.*

The contours after H and after LH with higher starting pitch tend to end at a similar mid range pitch height, and the contours after L and after HL with lower starting pitch tend to approximate to a similar lower pitch. The end point tone specification, H or L, of the preceding tone tends to correlate with the end point of the neutral tone. In Figure 5, two Gaussian distributions (μ= -0.11, σ^2= 0.25; μ= -0.23, σ^2= 0.20) are shown to present the distribution of f_0 at the end of contours. The observed means in Gaussian distribution are slightly different from the Least Squares Means computed in the LME models.

3.3. patterns of two consecutive neutral tones

Table 4 summarises the results of the LME models for the contours of two consecutive neutral tones. The preceding tone category leads to significant differences in average height, overall slope, and curvature of the contour of neutral tones.

The prototype of the contours of two consecutive neutral tones in four different contexts is shown in Figure 6, based on the Least Squares Means of *c* coefficients computed in the LME models. The shape of contours in the normalised time range from -1 to 0.2 resembles those in Figure 4, where

neutral tone contours converge at two pitch levels: one at about -0.05, 5% lower than the average pitch of a speaker, and the other about -0.17, 17% lower than the average pitch. Figure 6 further shows the four contours continue to converge at a low pitch at around -0.18.

Table 4: *Fixed effects in the linear mixed-effect models for aspects of pitch contour of two consecutive neutral tones in [X-N_1-N_2] trisyllabic utterances*

Shape	Predictor	df1	df2	F	*p*
HEIGHT	Tone X	3	172.98	5.04	**<.001**
SLOPE	Tone X	3	174.58	19.11	**<.001**
CURVATURE	Tone X	3	151.87	7.15	**<.01**

Model: Shape ~ Preceding Tone X Category + (1| Speaker).

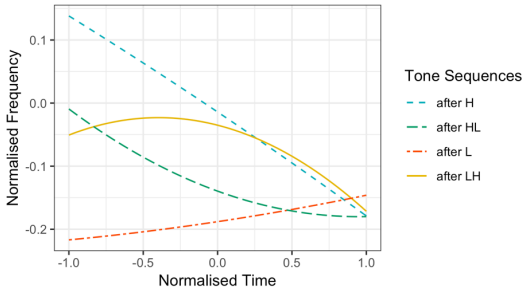

Figure 6: *Simulation of two consecutive neutral tone contours in different contexts.*

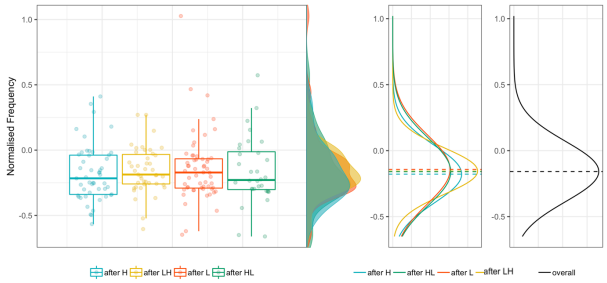

Figure 7: *Distribution of modelled ending pitch of the second neutral tone in [X-N1-N2] trisyllabic utterances.*

Figure 7 shows the distribution of normalised $f0$ at the end of the voicing part of the two consecutive neutral tones of all the utterances. The Gaussian distributions across the four tonal contexts are similar, as shown in Figure 7, where the distribution curves are superimposed on each other. A uniform Gaussian distribution (μ = -0.16, σ^2 = 0.23) was drawn based on all the end point f_0 data. Such convergence suggests that neutral tone has a pitch target and it is fairly low.

4. Discussion

Consistent with many previous studies, the duration of a neutral tone syllable is in general significantly shorter than its preceding lexical tone syllable. But in Table 1, we also find the neutral tone syllable men /mən/ is significantly longer than its preceding syllable. Similarly in the study [13] on Taiwan Mandarin, no evidence was found that the neutral tone syllable *men* is reduced in duration while neutral tone syllables *zi, zhe, de, le, ge* are reduced in duration. Such findings suggest syllable structure may play a role in the duration of the neutral tone because one major difference between *men* and the other neutral tones is that *men* has a slightly more complex structure of a closed syllable with a nasal coda.

One explanation proposed here is that unstressed neutral tone provides a context for vowel shortening in CV syllables, similar to closed syllables. The length of neutral tone syllable is sensitive to its syllable structure.

Neutral tones were also reported to be correlated with lexical frequency independent of other factors. More frequent neutral tones tend to be shorter in duration, lower in pitch, and weaker in intensity [14]. Productive functional morphemes, particularly *de* and *le*, the most frequently-used morphemes, are thus shorter than many other syllables.

The contours of two consecutive neutral tones (Figure 6) complete the patterns of a single neutral tone in disyllabic utterances that appear to have dual pitch targets (Figure 4). Instead of having varied pitch targets for neutral tone at different positions, the conception of target approximation [15] indeed offers a convincing way in explaining the findings. Different from the conclusion of [15] that Mandarin neutral tone has a underlyingly mid-level pitch target, however, our findings suggest a low level pitch target for neutral tones, although the pitch target for the converging trend at the end of two consecutive neutral tones seems slightly higher than the end pitch for L tones. Our results are similar to [13] that a static pitch target in the mid-low to low range was identified for Taiwan Mandarin.

Although an underlying L target can be assigned for neutral tones in our analysis, it cannot be neglected that the pitch realisation pattern of neutral tone is distinct from that of L tone where the f_0 usually reach a low pitch range by the end of the L tone. Neutral tone does not trigger the classic T3 tone sandhi phenomenon where the L tone is turned into a LH tone when it precedes another L tone. The differentiation between lexical L tone and neutral tone, that both have L pitch targets can be accounted for by the concept of articulation strength [16] in the soft template model [16, 17] or implementation strength in the PENTA model [15, 18]. In an intuitive sense, neutral tone tends to have much weaker strength and heavier carry-over influence from the preceding syllable. Such weakness in our concept can also be realised by a larger variance in a probabilistic model [19, 20] so that more variation can be tolerated. The variance distribution differentiates a lexical L tone from a neutral L tone, and that a lexical L tone has more specified and less variable f_0 realisation than a neutral L tone.

5. Acknowledgements

The author would like to thank prof. John Coleman and Jiahong Yuan for providing the Mandarin Hub4 Corpus and the corresponding time-aligned transcripts.

6. References

[1] S. Duanmu, "Syllable structure," In R. Sybesma (Ed.), *Encyclopedia of Chinese language and linguistics*, vol. 4, pp. 230-236, Leiden: Brill, 2017

[2] Y. R. Chao, *A grammar of spoken Chinese*. Berkeley and Los Angeles: University of California Press, 1968.

[3] M. Y. Chen, *Tone sandhi: Patterns across Chinese dialects.* Cambridge: Cambridge University Press, 2000.
[4] M. Yip, *Tone*. Cambridge: Cambridge University Press, 2002.
[5] J. Yuan, & M. Liberman, "Investigating consonant reduction in Mandarin Chinese with improved forced alignment," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2675–2678, 2015.
[6] J. W. Eaton, D. Bateman, S. Hauberg, & R. Wehbring, *GNU Octave version 4.0.0 manual: A high-level interactive language for numerical computations*. Retrieved from http://www.gnu.org/software/octave/doc/interpreter/. 2015.
[7] E. Grabe, G. Kochanski, & J. Coleman, "Quantitative modelling of intonational variation," *Speech Analysis and Recognition in Technology, Linguistics and Medicine*, pp. 45-57, 2003.
[8] E. Grabe, G. Kochanski, & J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," *Language and Speech,* vol. 50, no.2, pp. 281-310, 2007.
[9] R Core Team. *R: A language and environment for statistical computing.* Retrieved from https://www.R-project.org/. 2017.
[10] D.Bates, M. Maechler, B. Bolker, & S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software,* vol. 67,no.1, pp.1-48, 2015.
[11] A. Kuznetsova, P. B. Brockhoff, & R. H. Christensen, "lmerTest package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software,* vol. 82,no.13, pp. 1-26, 2017.
[12] F. Wu, & M. Kenstowicz, "Duration reflexes of syllable structure in Mandarin," *Lingua*, vol. 164, pp. 87–99. 2015.
[13] K. Huang, "Phonological identity of the neutral-tone syllables in Taiwan Mandarin: An acoustic study," *Acta Linguistica Asiatica,* vol. 8, no. 2, pp. 9-50, 2018.
[14] H. Kong, & S. Wu, "Frequency Effect and Neutralization of Tones in Mandarin Chinese," *Journal of Quantitative Linguistics*, vol. 26, no.2, 95–115, 2019.
[15] Y. Chen, & Y. Xu, "Production of weak elements in speech – Evidence from F_0 patterns of neutral tone in Standard Chinese," *Phonetica,* vol. 63, pp. 47-75, 2006.
[16] G. Kochanski, C. Shih, & H. Jing, "Hierarchical structure and word strength prediction of Mandarin prosody," *International Journal of Speech Technology,* vol. 6, pp. 33-43, 2003.
[17] G. P. Kochanski, & C. Shih, "Automatic modeling of Chinese intonation in continuous speech," *Proceedings of Eurospeech*, pp. 911-914, Aalborg: Eurospeech, 2001.
[18] Y. Xu, & S. Prom-on, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesising speech melody via model-based stochastic learning," *Speech Communication,* vol. 57, pp.181-208, 2014
[19] C. S. Blackburn, & S. Young, "A self-learning predictive model of articulator movements during speech production," *The Journal of the Acoustical Society of America*, vol.107, no.3, pp.1659-1670, 2000.
[20] J. Coleman, M. E. L. Renwick, & R. A. M. Temple, "Probabilistic underspecification in nasal place assimilation," *Phonology*, vol. 33, pp. 425-458, 2016.