# Measuring Gradient Effects of Alcohol on Speech with Neural Networks' Posterior Probability of Phonological Features

[1]Ratree Wayland, [1, 2]Kevin Tang, [3]Sophia Vellozzi, [1]Fenqi Wang and [3]Rahul Sengupta

[1]Department of Linguistics & [3]Department of Computer, Information Science and Engineering, University of Florida; [2]Department of English Language and Linguistics, Heinrich-Heine-Universität Düsseldorf
ratree@ufl.edu, kevin.tang@hhu.de, s.vellozzi@ufl.edu, fenqi@ufl.edu, rahulseng@ufl.edu

## ABSTRACT

Alcohol is known to impair fine articulatory control and movements. In drunken speech, incomplete closure of the vocal tract can result in deaffrication of the English affricate sounds /tʃ/ and /ʤ/, spirantization (fricative-like production) of the stop consonants and palatalization (retraction of place of articulation) of the alveolar fricative /s/ (produced as /ʃ/). Such categorical segmental errors have been well-reported. This study employs a phonologically-informed neural network approach to estimate degrees of deaffrication of /tʃ/ and /ʤ/, spirantization of /t/ and /d/ and place retraction for /s/ in a corpus of intoxicated English speech. Recurrent neural networks were trained to recognize relevant phonological features [anterior], [continuant] and [strident] in a control speech corpus. Their posterior probabilities were computed over the segments produced under intoxication. The results obtained revealed both categorical and gradient errors and, thus, suggested that this new approach could reliably quantify fine-grained errors in intoxicated speech.

**Keywords**: alcohol, deaffrication, palatalization, retraction, neural network.

## 1. INTRODUCTION

Alcohol intoxication has been shown to impair cognitive function and production of both suprasegmental and segmental properties of speech [1, 2, 3, 4, 5, 6, 7]. In English, the segmental errors encompass deaffrication of the affricate sounds /tʃ/ and /ʤ/ [5, 7], spirantization (fricative-like production) of the stop consonants [5, 7] and palatalization (retraction of place of articulation) of the alveolar fricative /s/ (produced as /ʃ/) [5, 9]. These errors occur due to impaired ability to control timing and movement of the active articulators, resulting in failure to form a complete closure (deaffrication and spirantization) or to achieve and/or maintain an appropriate degree of opening, at an intended location in the vocal tract (place retraction). Crucially, these errors have been described as categorical through such coarse-grained measures as perceptual judgment [e.g., 8] and phonetic transcription of the affected speech segments [e.g., 2]. However, sub-contrastive or gradient errors, below the level of a segment or feature are likely missed by such techniques due to the listeners' perceptual bias [10]. Further, while acoustic measurements could circumvent perceptual bias, describing categorical rather than gradient phonetic errors have largely been the focus of most acoustic-phonetic studies of intoxicated speech.

The goal of this study is to examine the nature, gradient and categorical errors, at the feature level in intoxicated speech in an English corpus using a neural network model known as Phonet. Inspired by computational approaches using forced alignment to measure surface, gradient phonetic variations, this approach quantifies gradient phonetic variation of deaffrication, spirantization and place retraction from the posterior probability of relevant phonological features, computed directly from the input signals by bidirectional recurrent neural networks. In this study, the relevant phonological features are [anterior], [continuant] and [strident]. These features capture relative location of the oral constriction, amount of airflow impedance and intensity of frication noise, respectively. A categorical error is operationalized as a sign shift of the phonological features (e.g., from [+anterior] to [-anterior]), while gradiency of an error is reflected in the posterior probability values of a phonological feature. A brief description of Phonet is present in § 2 below.

## 2. PHONET

Phonet [11] is a bi-directional recurrent neural network model, trained to recognize input phones as belonging to different phonological classes defined by phonological features (e.g., anterior, continuant and strident). It is semi-automatic and only requires a segmentally-aligned acoustic corpus (using forced alignment). Input to Phonet is log-energy distributed across triangular Mel filters computed from 25-ms windowed frames of each 0.5 second chunk of the input signal (see [11] for details). Once trained, posterior probabilities for different phonological features of the target segments can be computed by the model. Phonet has been found to be highly accurate in quantifying degree of lenition in Spanish [7, 15, 21, 22] and modelling the speech impairments of patients diagnosed with Parkinson's disease [11].

The architecture of Phonet is described in detail in [11].

Phonet can be customized with different sets of phonological features and acoustic representations. In this study, we focus on the probability of the phonological features [anterior], [strident] and [continuant] to capture retraction of place of articulation, deaffrication, and spirantization of English voiceless and voiced stops /t, d/, voiced and voiceless affricates /tʃ, dʒ/, and fricatives /s, ʃ/. A fricative-like realization due to an incomplete closure of the oral constriction would be associated with a relatively higher [continuant] and [strident] probability while a relatively low [anterior] probability would indicate a more retracted place of articulation.

## 3. METHODS

### 3.1. Materials

The target consonants for this study are English stops /t, d/ (Ns = 2,237 and 1,085), affricates /tʃ, dʒ/ (Ns = 144 and 160), and fricatives /s, ʃ/ (Ns = 1,430 and 96) from a corpus of intoxicated speech [7].

### 3.2. Stimulus recording procedure

The corpus contains recordings of four female, native speakers of British English reading a dialogue naturally (i.e., not in an animate, acting voice). The original text of the dialogue (based on [17]) was edited to ensure that it is gender- and emotionally neutral, void of overly long turns, and representative of the English phonemic inventory (available at [18]). Two separate readings (sober then drunk) across participants, on different days (1-2 months apart) were recorded in a sound-attenuated room at 44.1 kHz sampling rate and 16-bit amplitude resolution in stereo and were then converted to mono using Audacity. The speakers were told not to eat, drink, or use mouthwash 2 hours before each session and not to smoke half hour before each session.

The participants' blood alcohol concentration (BAC) was measured using a breathalyzer [AlcoMate (Macomb Township. MI) Premium AL-7000] at the beginning of the sober session to ensure absence of alcohol in their system. Intoxicated recording session began when BAC reached 0.12% after consumption of vodka or rum, mixed with juices.

### 3.2. Stimulus pre-processing

The recordings were divided up into utterances which were then manually annotated for any disfluencies. The disfluent utterances (8.5%) were not discarded since their exclusion did not qualitatively change our findings. The utterances were forced aligned using the Montreal Forced Aligner (version: 2.0) [12] with its released pretrained English model.

### 3.3. Phonet training procedure

Librispeech [16], a large corpus of English audiobooks was used as a representative English speech sample. A subset of the cleaned portion of 360 hours was selected. The corpus was then forced aligned using the Montreal Forced Aligner (version: 2.0) [12]. The phone set was set to IPA. For other parameters default values were used.

Model training with Phonet was performed on an NVIDIA GeForce RTX 3090 GPU using the Keras [19] library. The corpus was randomly split into a train subset (80%) and a test subset (20%) using the Python (Version 3.9) scikit-learn library [14]. Twenty-one Phonet models were trained for 20 phonological classes (consonantal, syllabic, voicing, labial, coronal, dorsal, lateral, nasal, rhotic, anterior, continuant, sonorant, strident, diphthong, high, low, back, round, stress, tense), and pause.

The model was highly accurate showing unweighted average recall (UAR) ranges from 91% (coronal)-98% (pause). Critically, the UARs for the anterior, continuant and strident features are 93%, 92% and 97%, respectively. The model was then applied to our selected word tokens from our forced-aligned intoxicated speech corpus with /t, d, tʃ, dʒ, s, ʃ/. The predictions were computed for 10-ms frames. For a token containing multiple frames, the average prediction from the middle frame(s) was taken as its prediction. Anterior, continuant and strident posterior probabilities obtained for each target consonant are then used for statistical analyses.

### 3.4. Statistical analyses

All statistical analyses were performed using the lme4 package [13] in R [13]. Contrast coding (-0.5, 0.5) was used for binary categorical variable. Random variables included speaker and word. Two complementary analyses were performed. First, to examine if posterior probabilities could predict drinking status, for each target consonant, a binary logistic regression analysis was performed with the three posterior probabilities (anterior, continuant and strident) as predictors and drinking status (sober and intoxicated) as dependent variable using the *glmer* function. A contrastive or categorical error was inferred when a feature emerged as the significant predictor. Second, to evaluate gradiency of an error, drinking status was entered as predictors in linear regression models (*lmer*) performed to investigate its predictions on the three posterior probability values for each target consonant. Increase or decrease in posterior probability of a feature indicate degrees of error gradiency. For both analyses, “drunk” was the reference level. Based on previous literature [5, 7], a higher continuant and strident posterior probability is expected for /t, d/ and /tʃ, dʒ/ while a lower anterior probability is expected for the fricative /s/ in the intoxicated relative to the sober condition [5, 9].

## 4. RESULTS

Results of the binary logistic regressions analyses with anterior, continuant and strident posterior probabilities as predictors and drinking status as the categorical, binary dependent variable are summarized in Table 1.

| Cons. | Predictor | Odds Ratios | *P* value |
|---|---|---|---|
| /t/ | Anterior | 1.11 | 0.615 |
| | **Continuant** | **0.43** | **<.001** |
| | Strident | 0.88 | 0.353 |
| /d/ | **Anterior** | **2.14** | **0.046** |
| | **Continuant** | **0.46** | **0.009** |
| | Strident | 1.07 | 0.760 |
| /tʃ/ | Anterior | 1.88 | 0.437 |
| | **Continuant** | **0.18** | **0.023** |
| | Strident | 1.98 | 0.641 |
| /dʒ/ | Anterior | 1.17 | 0.866 |
| | **Continuant** | **0.01** | **0.015** |
| | Strident | 1.77 | 0.706 |
| /s/ | Anterior | 2.97 | 0.471 |
| | Continuant | 0.62 | 0.473 |
| | Strident | 14.35 | 0.153 |
| /ʃ/ | Anterior | 0.80 | 0.784 |
| | Continuant | 0.82 | 0.847 |
| | Strident | 1.51 | 0.954 |

**Table 1:** Results of the binary logistic analyses.

From this table, we can see that continuant posterior probability emerged as the only significant predictor of drinking status for /t/, /tʃ/ and /dʒ/. Specifically, as the continuant probability increases, the likelihood that the speakers were sober decreases (odds ratios <0.5). For /d/, both continuant and anterior probabilities are the significant predictors. However, as expected, as the anterior probability increases, the likelihood of the sober status increases (odds ratios >0.5). No significant predictor was found for /s/ [odds ratios = 0.62-14.35, p>.005] and /ʃ/ [odds ratios = 0.80-1.51, p>0.05]. However, the >0.5 odd ratios indicated that as the anterior, continuant and strident probabilities increase, the likelihood of the sober status also increases. These results suggested that categorical errors (i.e., [-continuant] > [+continuant]) occurred under intoxication for /t/, /d/, /tʃ/ and /dʒ/. Additionally, for /d/, a categorical shift from [+anterior] > [-anterior] also occurred. On the other hand, no categorical error was detected for /s/ or /ʃ/.

Tables 3a, b and c summarize the results of the linear mixed-effect regression models with drinking status as predictors (reference level = drunk) and posterior probabilities of the three phonological features as the dependent variables.

The results obtained indicated that a significantly higher anterior probability for /t/, /d/ and /s/ [*β*s= 0.029, 0.042, 0.013; *p*s≤.001] is predicted for the sober speech relative to the drunken speech, but a non-significant change in anteriority between the two speech conditions is predicted for /tʃ/, /dʒ/, and /ʃ/ [*β*s =0.009, -0.037, -0.014; *p*s= >.0.05] (Table 2a). β values suggested that sober /tʃ/ is more anterior than drunk /tʃ/ while sober /dʒ/ and /ʃ/ are less anterior than drunk /dʒ/ and /ʃ/. In other words, a shift in place of articulation is significantly greater for /t/, /d/ and /s/ than for /tʃ/, /dʒ/, and /ʃ/ under intoxication.

| Predictor | Consonant | *β* | *P* value |
|---|---|---|---|
| **Anterior** | **/t/** | **0.03** | **0.001** |
| | **/d /** | **0.04** | **<0.001** |
| | /tʃ/ | 0.009 | 0.830 |
| | /dʒ/ | -0.037 | 0.399 |
| | **/s/** | **0.01** | **0.001** |
| | /ʃ/ | -0.014 | 0.746 |

**Table 2a:** Summary of the linear regression models for anterior probability.

For continuant probability (Table 2b), a significantly lower value is predicted for /t/, /d/, /tʃ/ and /dʒ/ [*β*s=-0.098, -0.061, -0.102, -0.068; *p*s≤.01], but not for /s/ [*β* = 0.009, *p*=0.098] or /ʃ/ [*β* = -0.005, *p* = 0.88] under the sober condition compared to the drunk condition. β value is positive for /s/, but negative for /ʃ/, indicating that /s/ is more continuant when sober while the opposite is true for /ʃ/. These results suggest that oral constriction for stops, and affricates became significantly less complete under intoxication. In contrast, oral constriction size was not significantly altered for the two [+continuant] consonant, /s/ and /ʃ/, when the speakers became inebriated.

| Predictor | Consonant | *β* | *P* value |
|---|---|---|---|
| **Continuant** | **/t/** | **-0.098** | **<0.001** |
| | **/d /** | **-0.061** | **<0.001** |
| | **/tʃ/** | **-0.103** | **0.013** |
| | **/dʒ/** | **-0.068** | **0.003** |
| | /s/ | 0.009 | 0.098 |
| | /ʃ/ | -0.005 | 0.880 |

**Table 2b:** Summary of the linear regression models for continuant probability.

Finally, Table 2c shows that strident probability for /t/ was predicted to be significantly lower under the sober condition while the opposite is true for /s/. No significant difference [*p*>.05] was predicted for the remaining consonants. Although statistically non-significant, β values for /d/ [-0.022] and /tʃ/ [-0.017] are negative suggesting less stridency in their production when sober than when drunk while that of /dʒ/ [0.022] is positive and that of /ʃ/ [0.000] equals to 0, suggesting more stridency for /dʒ/ but no change in stridency for /ʃ/ when sober. These results suggest that /t/ is significantly less strident (less turbulent noise) when produced under the sober condition. On the contrary, drunk /s/ is less strident than its sober version. Furthermore, minimal change in degrees of stridency is observed for /d/, /tʃ/ and /dʒ/ while no change is predicted for /ʃ/.

| Predictor | Consonant | *β* | *P* value |
|---|---|---|---|
| **Strident** | **/t/** | **-0.064** | **<0.001** |
| | /d / | -0.022 | 0.175 |
| | /tʃ/ | -0.017 | 0.492 |
| | /dʒ/ | 0.022 | 0.451 |
| | **/s/** | **0.014** | **0.001** |
| | /ʃ/ | 0.000 | 0.943 |

**Table 2c:** Summary of the linear regression models for strident probability.

## 5. DISCUSSION AND CONCLUSION

To detect categorical and gradient errors in intoxicated speech, a new computational approach, Phonet, was applied to a corpus of intoxicated English speech. Target error types are deaffrication, spirantization and retraction of place of articulation whose degrees of variation are estimated from posterior probabilities of three phonological features, [anterior], [continuant], [strident].

Binary logistic regression models yielded [continuant] as the significant predictor of drinking state for /t/, /tʃ/ and /dʒ/ while both [continuant] and [anterior] emerged as the significant predictors for /d/. If a binary, categorical shift of a feature is responsible for its significant predictive power, then, these results suggested that categorical errors ([-continuant] > [+continuant]) occurred for /t/, /d/, /tʃ/ and /dʒ/ under intoxication, suggesting that the size of the oral constriction contrastively shifts from a sober to a drunk state. For /d/, a categorical shift from [+anterior] > [-anterior] also occurred in drunken state, indicating that a concurrent and significant amount of place retraction also took place. The fact that /tʃ/ and /dʒ/ are [-anterior] may account for why further place retraction did not occur. In turns, neutralization (loss of contrastivity) in anteriority between /t/ and /tʃ/ could account for why /t/ did not undergo place retraction since they would both be [-anterior] if /t/ retracted. This suggest that articulatory planning may be intact, but the fine-grained motor control is partially lost when intoxicated.

Surprisingly, no categorical error was committed for /s/ or /ʃ/ under intoxication, at least not at the BAC level tested. Previous literature led to an expectation that a categorical shift in place of articulation would occur for /s/ (i.e., [+anterior] > [-anterior]) [5, 9]. It is possible that this error is only attested at a higher BAC level. However, it is worth noting that /s/ and /ʃ/ are both [+continuant, + strident]. It is possible that these shared and redundant features “add additional motoric instructions to enhance the saliency of the jeopardized features” [20, p. 33], namely [anterior] in this case. Nonetheless, the fact that this error has been previously reported suggested a limit of this enhancement effect.

Gradient errors are revealed by linear regression analyses. For instance, a shift in place of articulation was found to be significantly greater for /t/, /d/ and /s/ than for /tʃ/, /dʒ/, and /ʃ/ under intoxication. These results suggest that degree of place retraction in intoxicated speech is constrained by the existing place feature of the affected consonants: [+anterior] consonants will sustain a greater degree of place shift than [-anterior] consonants.

A similar constraint is observed for [continuant] leading to gradiency in continuant error. Continuant posterior probability significantly increased for [-continuant] consonants, /t/, /d/, /tʃ/, and /dʒ/, but not for [+continuant] consonant, /s/ and /ʃ/. This is due, perhaps, to a higher degree of continuance (greater oral aperture) which would lead to reduction of intensity of frication noise (i.e., stridency).

Finally, gradient errors involving [strident] was also found. Sober /t/ is significantly less strident than drunk /t/, suggesting an incomplete closure resulting in stridency. In contrast, a change in stridency was relatively small for /d/, /tʃ/ and /dʒ/. These results suggest that size of oral opening of [-continuant] consonants could increase. In addition, drunk /s/ is significantly less strident than its sober variant, suggesting a further widening of the oral aperture leading to a loss in stridency. However, no change was observed for /ʃ/. This result suggests that oral aperture could further widen for [-anterior, +continuant, +strident], /s/, but not for [+anterior, +continuant, +strident], /ʃ/.

Both categorical and gradient errors are revealed by Phonet, suggesting that it could reliably quantify fine-grained errors in intoxicated speech. Our findings need to be confirmed with more subjects, and can be extended to languages with different contrastive phonological features [7], and compared with speech by clinical populations, such as Parkinson's disease.

**Acknowledgement:** This research was supported by NSF-National Science Foundation (SenSE). Grant No. 2037266. We acknowledge the data contributions of Jeremy Grynpas and Andrew Nevins.

## 6. REFERENCES

[1] Arbuckle, T. Y., Chaikelson, J. S., Gold, D. P. 1994. Social drinking and cognitive functioning revisited: the role of intellectual endowment and psychological distress. *Journal of studies on alcohol*, 55(3), 352-361.

[2] Pisoni, D. B., Martin, C. S. 1989. Effects of alcohol on the acoustic-phonetic properties of speech: perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research*, 13(4), 577-587.

[3] Pisoni, D. B., Johnson, K., Bernacki, R. H. 1991, September. Effects of alcohol on speech. *In Proc. of the Human Factors Society Annual Meeting* (Vol. 35, No. 10, pp. 694-698). Sage CA: Los Angeles, CA: SAGE Publications.

[4] Swartz, B. L. 1992. Resistance of voice onset time variability to intoxication. *Perceptual and motor skills*, 75(2), 415-424.

[5] Lester, L.; Skousen, R. 1974. The Phonology of Drunkenness. In: Bruck, A.; Fox, RA.; LaGaly, MW. (eds.), *Papers from the Parasession on Natural Phonology*. Chicago: Chicago Linguistic Society.

[6] Hollien, H., Dejong, G., Martin, C. A., Schwartz, R., Liljegren, K. 2001. Effects of ethanol intoxication on speech suprasegmentals. *J. Acoust. Soc. Am*., 110(6), 3198-3206.

[7] Tang, K., Chang, C. B., Green, S., Bao, K. X., Hindley, M., Kim, Y. S., & Nevins, A. 2022. Intoxication and pitch control in tonal and non-tonal language speakers. *JASA Express Letters*, 2(6), 065202.

[8] Zihlmann, U. 2017, August. The Effects of Real and Placebo Alcohol on Deaffrication. In *Interspeech* (pp. 3882-3886).

[9] Johnson, K., Pisoni, D. B., Bernacki, R. H. 1990. Do voice recordings reveal whether a person is intoxicated? A case study. *Phonetica*, 47(3-4), 215-237.

[10] Frisch, S.A., Wright, R. 2002. The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, 30(2), 139-162.

[11] Vásquez-Correa, J. C., Klumpp, P., Orozco-Arroyave, J. R., Nöth, E. 2019. Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech. In *Interspeech*, 549-553.

[12] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2017. August. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*, 498-502.

[13] R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 969.

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... , Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825-2830.

[15] Wayland, R., Tang, K., Wang, F., Vellozzi, S., and Sengupta, R. 2022. Lenition measures: Neural networks' posterior probability vs. acoustic cues. *Proc. of Meetings on Acoustics*, vol. 50, no. 1, p. 060002.

[16] Panayotov, V., Chen, G., Povey, D., Khudanpur, S. 2015, April. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206-5210. IEEE.

[17] Simon, N. 1986. *The Collected Plays of Neil Simon: Volume 2*. New York, NY: Plume.

[18] Tang, K., Chang, C. B., Green, S., Bao, K. X., Hindley, M., Kim, Y. S., & Nevins, A. 2022, "Materials for Tang et al. (2022)" Available: osf.io/y2r87.

[19] Chollet, F., others. 2015. Keras. GitHub. Retrieved from https://github.com/fchollet/keras.

[20] Keyser, S. J., Stevens, K. N. 2006. Enhancement and overlap in the speech chain. *Language*, 33-63.

[21] Tang, K., Wayland, R., Wang, F., Vellozzi, S., Sengupta, R., & Altmann, L. 2023. From sonority hierarchy to posterior probability as a measure of lenition: The case of Spanish stops. *The Journal of the Acoustical Society of America*, 153(2), 1191-1203.

[22] Wayland, R., Tang, K., Wang, F., Vellozzi, S., & Sengupta, R. (2023). Quantitative Acoustic versus Deep Learning Metrics of Lenition. Languages, 8(2), 98.