SPEAKER STATISTICAL AVERAGENESS MODULATES WORD RECOGNITION IN ADVERSE LISTENING CONDITIONS

William L. Schuerman[1,2], James M. McQueen[3,4], Antje Meyer[3,4]

1. Department of Neurological Surgery, UCSF 2. Weill Institute for Neurosciences, UCSF 3. Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University 4. Max Planck Institute for Psycholinguistics

William.Schuerman@ucsf.edu

ABSTRACT

We tested whether statistical averageness (SA) at the level of the individual speaker could predict a speaker's intelligibility. 28 female and 21 male speakers of Dutch were recorded producing 336 sentences, each containing two target nouns. Recordings were compared to those of all other same-sex speakers using dynamic time warping (DTW). For each sentence, the DTW distance constituted a metric of phonetic distance from one speaker to all other speakers. SA comprised the average of these distances. Later, the same participants performed a word recognition task on the target nouns in the same sentences, under three degraded listening conditions. In all three conditions, accuracy increased with SA. This held even when participants listened to their own utterances. These findings suggest that listeners process speech with respect to the statistical properties of the language spoken in their community, rather than using their own speech as a reference.

Keywords: Word recognition, SA, speech in noise, noise-vocoded speech, dynamic time warping

1. INTRODUCTION

Even within a small linguistic community, speech characteristics can vary greatly within and between speakers. One noticeable dimension along which speech can vary is its *intelligibility*. Individuals possess idiosyncratic speech styles that can vary greatly in intelligibility. Yet there seems to be much less variability in *perception*. Listeners label certain speakers as more intelligible or less intelligible with high inter-rater agreement [4]. Accordingly, in everyday conversation we need not specify that a specific speaker is particularly intelligible *to a specific listener*. Instead, we simply say that a speaker is "intelligible", tacitly assuming this to mean that that speaker will be – barring complications such as hearing loss – intelligible *to all listeners*. How is it that individuals with idiosyncratic speech styles converge during perception? Despite the wealth of research devoted to analyzing the statistical properties of speech sounds as well as listeners' sensitivity to such statistics (e.g. [8], [3]), rarely do we consider Statistical Averageness (SA) of individual speakers as a property that could modulate intelligibility.

In this study, we examined whether speakers whose speech was more aligned to the statistical average of their linguistic community would be more intelligible under adverse listening conditions. In the first of four sessions, we recorded native speakers of Dutch producing short sentences, each containing two target nouns. Using *dynamic time warping*, we compared productions of the same sentence across multiple speakers to quantify the degree to which an individual's productions differed from those of all others in the cohort (i.e., how statistically average that individual was). In three subsequent sessions, the same participants attempted to recognize the target words under three types of degradation: 1) Noise-Vocoded Speech (NVS), which can eliminate fine-grained spectral cues to speaker identity; 2) Speech-in-Noise (SPIN), which preserves such cues; 3) Speech-in-Noise that had been filtered to approximate how one's own speech sounds during production (FiltSPIN). Our analysis examined whether variation in word recognition could be explained by a speaker's SA score.

We included the FiltSPIN session because the design of our study enabled us to examine an additional, related question: If a listener has a lower SA score, does this also mean that intelligibility is diminished when they are attempting to recognize self-produced words, or does their extensive experience with the statistics of their own speech mitigate this effect? If so, this may depend on the spectral properties of the incoming speech matching what they normally hear during production.

For all sessions, we predicted that word recognition accuracy would be higher for items produced by speakers with higher SA scores (and vice versa). Positive results would indicate 1) that despite differing in their manner of production, individuals

share common representations during listening, and 2) that these representations reflect the statistical distributions of speech in their linguistic community.

2. METHODS

2.1. Participants

Forty-nine native speakers of Dutch (21=M) were recruited. All gave written consent prior to each session. Due to dropout over time, the NVS, SPIN and FiltSPIN word recognition sessions consisted of 46, 42, and 43 participants, respectively.

2.2. Lexical Stimuli and Recording

Each stimulus sentence was of the form "the WORD1 is {above, under, next to} the WORD2". In order to ensure that words of varying difficulty were equally allocated across speakers, we collected 112 Dutch words and sorted them into four Word Groups defined by lexical frequency and phonological neighborhood density.

Participants were seated in a sound-attenuated booth 3 to 5 centimeters from a pop-filter shielded microphone (Sennheiser ME 64). To elicit the target sentences without having participants simply read words off a screen, we developed a 'semi-spontaneous' elicitation method. Participants were presented with two words in succession (e.g., "witch – ball"), followed by a simple display in which two images (corresponding to the previously presented nouns) were arranged in a particular spatial configuration. Based on the order of the nouns and the configuration of the images, there existed only one "correct" sentence to produce (e.g.,"the witch is under the ball"). Therefore, the exact form of the sentence was pre-specified on each recording trial but at no time did participants simply read words off a screen. Participants were instructed to try to speak as naturally as possible. Each of the 112 target nouns appeared in both initial and final position, for a total of 112 sentences. Three lists were generated and participants were presented with all three lists, comprising 336 total sentences. A researcher monitored the recording session for errors.

2.3. SA Metric

In order to compute the SA metric for each speaker, we first compared recordings using dynamic time warping (DTW). DTW is an algorithm that attempts to find an optimal alignment between two vectors or matrices [11]. The cost function of this algorithm can be used as a metric to quantify the similarity between two sound files [16]. For each sound file we extracted 26 mel-frequency-cepstral-coefficients (MFCCs), which are commonly used for speech recognition programs, using the *librosa* Python package [9] to create an MFCC spectrogram matrix. For each sentence, we computed the DTW distance [13] between the sound files for that sentence for every combination of same-sex speakers, creating a distance matrix (336 sentences per pair). Averaging over sentences resulted in a single distance score between each pair of speakers. For each speaker, we averaged over all distance scores to obtain their average distance to each speaker. We then standardized the scores by subtracting the mean and dividing by the standard deviation. Finally, we took the reciprocal of the standardized scores so that a higher score would indicate that the speaker was more statistically average.

2.4. Degraded Stimuli Preparation

2.4.1. Noise Vocoding

Noise-vocoding can systematically degrade the spectral content of speech while preserving temporal properties [15]. We separated the acoustic spectrogram into six frequency bands. At each point in time, the average amplitude of all frequencies within a given band was extracted and these values were then utilized to modulate the energy of broad-spectrum noise in the corresponding frequency bands.

2.4.2. Speech in Noise

Speech in noise (SPIN) is a standard technique in which a speech signal is embedded in speech shaped noise at a specified signal-to-noise ratio (SNR). For this study, we utilized an SNR of -7 decibels.

2.4.3. Filtered Speech in Noise

Prior to embedding in noise, we applied a filter [17] that approximated the effects of bone-conduction and other processes that affect auditory feedback during production. This filter was intended to simulate how a person hears their own voice when they are actively speaking.

2.5. Word Recognition

Participants were assigned to groups of seven same-gender speakers; each participant was presented with an equal number of sentences produced by each speaker in their group (including the listeners them-

selves). Each word recognition session was defined by the type of degraded stimulus presented (NVS, SPIN, filtSPIN). The first session was always NVS. The order in which participants completed SPIN and filtSPIN was randomized across participants. Time between session 1 and session 2 averaged 7.4 days ($sd = 2.9$). Time between session 2 and session 3 averaged 14.7 days ($sd = 4$).

In each recognition session, participants were presented with manipulated versions of the recordings and asked to identify the two target nouns in each sentence by typing in their responses via a computer keyboard. In each session one speaker provided 48 sentences (96 target words), with 12 words from each word group in first position and 12 words from each word group in second position. This ensured that all sentences were presented without repetition, while balancing lexical factors across speakers.

In order to compare the typed-in responses to the auditorily presented target words, all target words and participant responses were broadly transcribed into DISC, a computer readable phonetic orthography. This eliminated the influence of orthographic variation with no phonetic realization (e.g., in Dutch, word final "t" and "d" are both realized as [t]). A response was marked as correct if the response and target transcriptions were identical.

3. RESULTS

Fig. 1 displays average accuracy for each participant as a function of SA, separated by Position (First, Second) and Session (NVS, SPIN, FiltSPIN). Given that by-item difficulty had been balanced across subjects in the experimental design, we elected to average over target word items and analyze the *proportion* of correct responses using linear mixed effects regression [1]. Model selection was carried out by backwards-fitting using maximum likelihood comparison until all non-significant terms had been removed.

Model comparison began with a full model containing fixed effects for Session, Word Position (First, Second), and SA, as well as all interaction terms, random intercepts for Subject, and random slopes over Subject for Session, Word Position, and SA (in R syntax: *Prop. Correct 1 + Session*Word Position*SA + (1+Session+Word Position+SA|Subject)*). All non-significant terms were removed until reaching a final model containing main effects for Session, SA, and Word Position, as well as an interaction term between Session and Word Position. Conditional R^2 (variance explained by both fixed and random effects) was 0.625, marginal R^2 (variance explained by fixed effects alone) was 0.43 [12]. Model estimates are reported in Table 1, with term-specific p values obtained via Satterthwaite's Method.

Table 1: Estimates from final model reported using treatment coding. Intercept represents average proportion of correct responses in first position of the NVS session.

Fixed Effect	Est.	Std. Error	Pr(>|t|)
Intercept	0.529	0.012	<0.001
Second Position	0.03	0.01	0.003
Session SPIN	0.152	0.012	<0.001
Session FiltSPIN	0.102	0.013	<0.001
SA	0.036	0.005	<0.001
Sec. Pos.:SPIN	-0.285	0.012	<0.001
Sec. Pos.:FiltSPIN	-0.294	0.012	<0.001

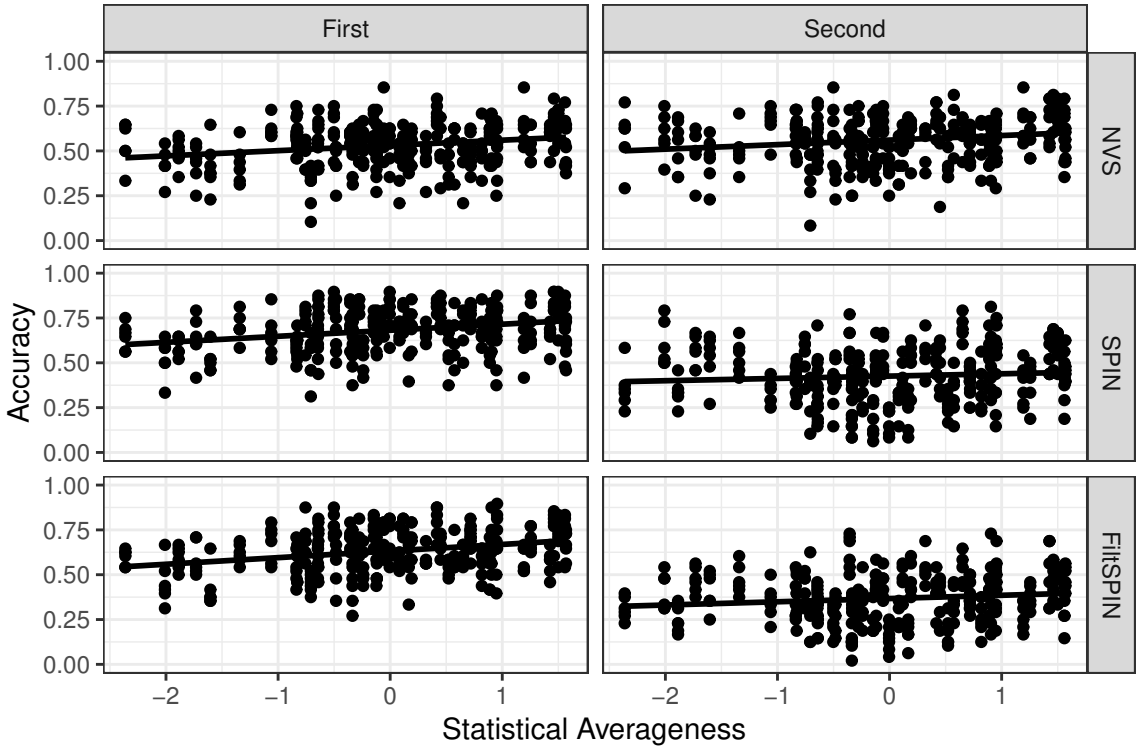
For target words appearing first in the sentence, accuracy was higher in the SPIN and filtSPIN sessions than in NVS. However, accuracy for words appearing in second position was much lower in these sessions than in NVS. In the NVS session, there is a positive relationship between accuracy and SA score for words in both positions. While the visual representation of the data in Fig. 1 suggests that in the SPIN and filtSPIN sessions the effect of SA was strongest for words in the first position, no significant interaction term was found between Session and SA.

Following up on this main analysis, we restricted the data to trials in which participants were listening to their own voice and re-ran the same model in order to determine whether the effect of SA was diminished when listening to one's own voice. This was not the case: the effect of SA was actually stronger (β=0.057, $p < 0.001$). Furthermore, no significant interaction was found with session, indicating that these effects did not differ significantly in the FiltSPIN condition (when the stimuli had been filtered to approximate the sound of the participant's own voice during speaking).

4. DISCUSSION

In this study, we developed a metric quantifying the statistical averageness (SA) of a speaker's speech with respect to a cohort of other speakers. In three separate word recognition tasks using different types of word degradation, we found a positive relationship between SA scores and accuracy. The higher a speaker's SA score, the more likely it was that listeners would be able to accurately recognize their

Figure 1: Accuracy by SA score, organized across rows by Session and columns by Position. The solid line in each plot represents best linear fit to data points.



speech under degraded conditions. This agrees with previous studies that have found that statistical averageness facilitates perception [14, 18].

However, in order to reduce experimental complexity, listeners only listened to same-sex talkers and SA was only computed with reference to same-sex talkers. Thus, we can only speculate as to how our results generalize beyond this somewhat constrained cohort. For example, given that the perceived gender of a talker has been found to influence speech perception [6], it may be that statistical averageness of a talker is perceived with reference to the perceived sex of the talker.

The claims of this study rest heavily on the way in which the SA metric was computed. The algorithm we decided to employ, using MFCCs and dynamic time warping, was relatively uninformed and made few assumptions about the structure of the data. One advantage was that the algorithm could be applied to the entire sound file, without the need to extract specific speech segments (e.g., consonants and vowels). A disadvantage was that the method treated all MFCCs as equally informative. It may be the case that certain MFCCs only contributed noise to the SA metric or that other phonetic features could be more informative than MFCCs. Methodological comparisons suggest that extracting important phonetic features from recordings may yield better results [7]. Based on such comparisons, we would predict the effects of SA to be even stronger when computed using a more informative metric.

Our results indicated that statistically average speakers were more intelligible in difficult listening conditions than less typical speakers. This was true even when participants were listening to recordings of their own voices. These results are problematic for theories suggesting that listeners interpret incoming speech by reference to representations or simulations based on their production experience [10], especially when speech is degraded [2]. Instead, the results accord with models of speech perception which place primary emphasis on auditory experience [5]. Our findings indicate that listeners process incoming speech with respect to the statistical properties of the speech of their linguistic community.

5. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* 67(1).

[2] D'Ausilio, A., Bufalari, I., Salmas, P., Fadiga, L. 2012. The role of the motor system in discriminating normal and degraded speech sounds. *Cortex* 48(7), 882–887.

[3] Feldman, N. H., Griffiths, T. L., Morgan, J. L. 2009. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review* 116(4), 752–782.

[4] Hazan, V., Markham, D. 2004. Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America* 116(5), 3108–3118.

[5] Holt, L. L., Lotto, A. J. 2008. Speech Perception Within an Auditory Cognitive Science Framework. *Current Directions in Psychological Science* 17(1), 42–46.

[6] Johnson, K., Strand, E. A., D'Imperio, M. 1999. Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics* 27(4), 359–384.

[7] Keen, S., Ross, J. C., Griffiths, E. T., Lanzone, M., Farnsworth, A. 2014. A comparison of similarity-based approaches in the classification of flight calls of four species of North American wood-warblers (Parulidae). *Ecological Informatics* 21, 25–33.

[8] Kuhl, P. K. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics* 50(2), 93–107.

[9] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., Nieto, O. 2015. Librosa: Audio and Music Signal Analysis in Python. 8.

[10] Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., Iacoboni, M. 2007. The Essential Role of Premotor Cortex in Speech Perception. *Current Biology* 17(19), 1692–1696.

[11] Müller, M. 2007. Dynamic Time Warping. In: *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer Berlin Heidelberg 69–84.

[12] Nakagawa, S., Schielzeth, H. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2), 133–142.

[13] Rouanet, P. 2018. Dtw: DTW (Dynamic Time Warping) python module.

[14] Schuerman, W. L., Meyer, A., McQueen, J. M. 2015. Do We Perceive Others Better than Ourselves? A Perceptual Benefit for Noise-Vocoded Speech Produced by an Average Speaker. *PLOS ONE* 10(7), e0129731.

[15] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M. 1995. Speech Recognition with Primarily Temporal Cues. *Science* 270(5234), 303–304.

[16] Somervuo, P. 2018. Time–frequency warping of spectrograms applied to bird sound analyses. *Bioacoustics* 0(0), 1–12.

[17] Vurma, A. 2014. The timbre of the voice as perceived by the singer him-/herself. *Logopedics Phoniatrics Vocology* 39(1), 1–10.

[18] Wöllner, C., Deconinck, F. J. A., Parkinson, J., Hove, M. J., Keller, P. E. 2012. The perception of prototypical motion: Synchronization is enhanced with quantitatively morphed gestures of musical conductors. *J Exp Psychol Hum Percept Perform* 38(6), 1390–1403.