PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Individual Articulatory Control in Speech Enrichment

Chen SHEN[1]; Martin COOKE[2]; Esther JANSE[3]

[1] Radboud University, the Netherlands

[2] Ikerbasque (Basque Science Foundation), Spain

[3] Radboud University, the Netherlands

ABSTRACT

Individual speakers may use various strategies to enrich their speech while speaking in noise (i.e., Lombard speech) to improve intelligibility. The resulting acoustic-phonetic changes in Lombard speech are highly variable amongst different speakers, but it is unclear what causes these talker differences. This study investigates the potential role of articulatory control in speakers' Lombard speech enrichment success. Individuals' predicted intelligibility in both speaking styles (presented at -5 dB SNR) was measured using a glimpse-based metric.

Seventy-eight speakers read out sentences in both their habitual style, and in a condition where they were instructed to speak clearly while hearing loud speech-shaped noise. A diadochokinetic speech task, that requires speakers to repetitively produce word or non-word sequences as accurately and as rapidly as possible, was used to quantify their articulatory control.

Speakers' predicted intelligibility scores were significantly higher in the Lombard than habitual speech condition. There was no simple effect of articulatory control on predicted intelligibility, nor an interaction between speaking condition and articulatory control. However, those with poorer articulatory control required some practice to maximise their enrichment success. This suggests that articulatory control may relate to a speaker's knowledge about the articulatory configuration that will provide the best intelligibility.

Keywords: Speech, Intelligibility, Articulation

1. INTRODUCTION

Speech communication rarely happens in noise-free conditions. Individual speakers often, consciously or unconsciously, use various strategies to enrich their speech in order to improve the intelligibility of their speech when talking in noisy environments (1). Speech produced in noisy conditions is known as 'Lombard Speech' (2), and the acoustic-phonetic changes that occur in Lombard speech are highly variable amongst different speakers (3). However, what causes these talker differences, and what impact these differences have on intelligibility, remains unclear. This study investigates the potential role of articulatory control in speakers' Lombard speech enrichment success, i.e., the degree to which they can enhance their intelligibility in noise, relative to the intelligibility of their habitual speaking style.

Articulatory control is often referred to as "the systems and strategies that regulate the production of speech, including the planning and preparation of movements and the executive of movement plans to result in muscle contraction and structural displacement" (4). Articulatory control in clinical settings is often measured through maximum performance speech tasks, such as an oral diadochokinetic (DDK) task. Despite the debates around whether DDK task is representative enough for speakers' speech motor control abilities (5), DDK remains one of the most widely used tasks in the assessment of speech motor disorders. The task is relatively easy to conduct and administer (6,7), and it specifically tests speakers' speech motor limitations (8).

One way to obtain intelligibility scores for speech (often in the presence of noise) is through collecting subjective ratings from normal-hearing or sometimes hearing-impaired listeners. However,

[1] c.shen@let.ru.nl
[2] m.cooke@ikerbasque.org
[3] e.janse@let.ru.nl

the process of acquiring subjective intelligibility ratings is time-consuming and resource-demanding, particularly when assessing speech materials of many speakers. To replace human (subjective) intelligibility ratings, several objective intelligibility measures have been proposed to predict speech intelligibility in the presence of background noise. For instance, in earlier studies, the articulation index (9,10), and the speech-transmission index (11) were the most commonly used metrics. More recently, a glimpse-based speech perception model was proposed (12). Through an internal automatic speech recognition component, the model recognises the speech-dominant spectro-temporal regions, or "glimpses", in speech that survives energetic (noise) masking, attempting to model human speech perception in noise (12). Subsequently, several studies have used the output of the initial stage of the glimpsing model, the amount of supra-threshold target speech surviving energetic masking or "glimpse proportion", as a proxy for intelligibility, aiming to predict speech intelligibility in a more economical way (e.g., 13,14).

Four glimpse-based metrics were recently evaluated to compare their capability of accounting for the subjective intelligibility of a variety of speech styles in a range of masker types (15). Of the four, the high-energy glimpse proportion (HEGP) metric had the highest correlations with human listeners' judgements across the tested datasets. Given our interest in investigating whether individual' articulatory control ability predicts speakers' Lombard speech enrichment success, and due to the large number of speakers (N = 78) to evaluate, we opted for the more robust metric (HEGP) for the current study. In addition, we investigated whether speakers' predicted intelligibility in the two speaking conditions changed over the course of the list of sentences they were asked to produce.

2. METHODOLOGY

2.1 Data collection

Seventy-eight speakers were recruited and recorded at the Centre for Language Studies lab at Radboud University. They were all native Dutch speakers, with no speech, hearing, or reading disabilities, nor past diagnosis of speech pathology or brain injury, and they all had normal or corrected-to-normal vision. Participants were reimbursed for their time through course credits or gift vouchers, and all 78 of them gave informed consent for their audio recordings to be analysed.

A DDK speech task was used to elicit participants' maximum performance (rate and accuracy) as an index of their articulatory control. Participants were instructed to repetitively produce two non-words ('pataka', 'katapa') and two real (Dutch) words ('pakketten' – *packages*, 'kapotte' – *broken*) as accurately and as rapidly as possible for around 10 seconds.

A sentence-reading task was used to elicit participants' habitual and Lombard speech. Participants were first instructed to read out 48 sentence stimuli correctly and fluently in their habitual style. Different lists were made to counterbalance the order of the 48 sentences over participants. They were then asked to read out the same 48 sentences (but in different orders) while being exposed to speech-shaped noise (at 78 dB SPL) through a pair of closed headphones (Sennheiser HD 215). In the latter condition, participants were also told to speak clearly and that their speech would be evaluated at a later stage, and that the top-three most intelligible speakers would get a gift voucher.

Stimuli of the two speech tasks were presented using PowerPoint slides on a 24'' full HD monitor placed in front of the participant. Recordings were made using a Sennheiser ME 64 cardioid capsule microphone through a pre-amplifier (Audi Ton) onto a steady-state 2 wave/mp3 recorder Roland R-05 in a sound-attenuating recording booth. The first author monitored participants' task progress and controlled the changing of stimuli slides outside the recording booth on the stimulus computer.

2.2 Articulatory Control

Participants' maximum performance (rate and accuracy) in the DDK speech task was used as an index of their articulatory control. Individual DDK articulation rate and accuracy were analysed acoustically in *Praat* (16), based on the first 7-second (or as close to 7-second as possible for the repetition counts to be an integer) of their DDK production. Rate (syllables/sec) was calculated by multiplying the total number of correct-and-full repetitions of (non)words produced by each participant in the 7-second time window by three (syllables), and divided this number of total syllables by the actual production time (total-duration minus errors, in-breaths, and pauses longer than 200 ms between repetitions). Accuracy (fraction) was calculated as number of correct repetitions divided by number of all repetitions in the same 7-second time window. Repetitions were only counted as correct if they were correct repetitions without any pauses longer than 200 ms.

2.3 HEGP-model predicted intelligibility

HEGP-metric predicted intelligibility scores were obtained per sentence (for the 48 sentences) in both habitual and Lombard speaking conditions for all speakers at -5 dB SNR. The speech-shaped noise employed in Lombard speech elicitation was used as the added noise masker for the HEGP calculation. HEGP scores were calculated based on the contribution of the high-energy glimpses to intelligibility surviving energetic masking. HEGP scores lie between 0 and 1, with higher numbers indicating higher glimpse proportions escaping energetic masking, thus higher predicted intelligibility.

2.4 Statistical analysis

To investigate whether speakers' Lombard speech enrichment success can be predicted by their articulatory control ability, we fitted two linear mixed-effects models using HEGP model-predicted intelligibility scores as dependent variable, one taking DDK rate (z-transformed), and the other taking DDK accuracy (z-transformed) as fixed effect of interest. Speech condition (habitual versus Lombard), and Trial number (the order in which a particular sentence appeared in the reading task) were also included as fixed factors in these models, as well as their potential interactions with DDK performance. Additionally, Participant and Sentence were included as random effects in the two models. Model fit improved for both models by allowing a random by-participant slope for the Condition effect, indicating that speakers indeed differed in their enrichment effect. The two full models (testing for a triple interaction between Condition, Trial and DDK performance) were then stripped in a step-wise manner (with the insignificant interactions removed first, followed by insignificant effects, starting from the ones that had the lowest t-values) to arrive at the most parsimonious model. Model comparisons were applied after each removal of the least significant predictor to verify that the exclusion of the predictor (or interaction) did not lead to significantly different model fit.

3. RESULTS

Table 1 presents descriptive data of speakers' mean performance in the DDK task. Speakers' mean intelligibility scores, as predicted by the HEGP metric, are 0.44 (*SD* = 0.02) for habitual speech and 0.54 (*SD* = 0.05) for Lombard speech respectively.

Table 1 – Summary of task performance

Speech tasks	Rate (syllables/sec)		Accuracy (fraction correct)	
	Mean	SD	Mean	SD
DDK	6.12	0.82	0.91	0.08

The overall Lombard intelligibility gain can also be seen in Figure 1 below, showing the averaged HEGP scores for each speaker (each dot represents one speaker's HEGP scores collapsed over the 48 sentences) in both habitual (x-axis) and Lombard (y-axis) conditions.

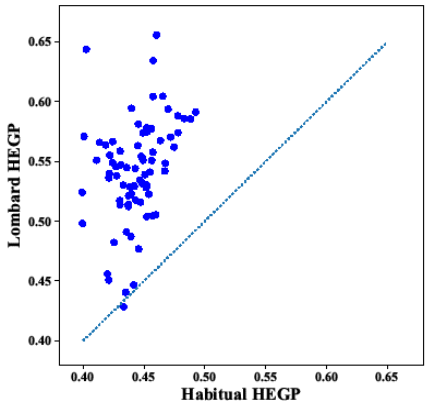

Figure 1 – Speakers' averaged HEGP scores in both speech styles. Note that values above the diagonal line represent a Lombard intelligibility gain.

Figure 1 illustrates the variation in intelligibility across speakers, particularly in the Lombard condition. As mentioned before, the two statistical models on the predicted intelligibility of the speakers tested whether DDK performance (DDK rate or DDK accuracy) modulated speakers' enrichment success (as quantified by the difference in predicted intelligibility between the two speech conditions), and whether speakers' predicted intelligibility changed over trials. DDK rate did not predict HEGP intelligibility, nor did it interact with Condition or Trial. However, the model of DDK accuracy (rather than DDK rate) showed the following results.

Table 2 – Linear mixed-effects model output for predicted speech intelligibility

	HEGP Scores		
Predictors	*Estimates*	*Std. Error*	*p*
(Intercept: Lombard)	0.54	0.59×10^{-2}	**<0.001**
Habitual	-0.10	0.47×10^{-2}	**<0.001**
DDK Accuracy	-0.23×10^{-2}	0.49×10^{-2}	0.631
Trial	0.17×10^{-3}	0.40×10^{-4}	**<0.001**
ConditionHabitual × DDK Accuracy	0.34×10^{-2}	0.47×10^{-2}	0.471
ConditionHabitual × Trial	-0.15×10^{-3}	0.48×10^{-4}	**0.002**
DDK Accuracy × Trial	-0.87×10^{-4}	0.27×10^{-4}	**0.001**
ConditionHabitual × DDK Accuracy × Trial	0.11×10^{-3}	0.39×10^{-4}	**0.006**

Table 2 confirms that predicted intelligibility in the habitual condition was lower than in the Lombard condition, and that predicted intelligibility in the Lombard condition increased over trials while it remained stable in the habitual condition (as shown in a model in which the habitual speaking condition was mapped on the intercept). DDK accuracy did not have an overall effect on predicted intelligibility, but it did interact with the increase in Lombard intelligibility over trials. More specifically, the intelligibility increase over trials, as observed in the Lombard condition, was smaller for those with higher DDK accuracy.

4. DISCUSSION

This study was set up to investigate the potential role of articulatory control in the extent to which speakers can enrich their speech when changing from their habitual speaking style to speaking clearly in noise (i.e., their Lombard speech enrichment success). Rather than collecting human (subjective) intelligibility ratings for our corpus of speakers, we obtained acoustic predictions for the intelligibility of our 78 speakers' habitual and Lombard speaking style. As expected, speakers indeed differed in their speech enrichment success. Furthermore, speakers generally needed some practice to arrive at slightly greater Lombard gains as evident from the trial effect in the Lombard speaking style. It is not the case that participants with better articulatory control (as indexed by DDK accuracy) were generally more intelligible, in either speaking condition. However, the ones with better DDK accuracy showed less improvement of their Lombard intelligibility through practice (over trials). In that sense, DDK does index speakers' articulatory control, if articulatory control is seen as speakers' knowledge of what to do articulatorily to produce more intelligible speech in noise. Further acoustic analyses are needed to examine whether speakers with poorer articulatory control changed from applying less appropriate to more appropriate enrichment strategies.

ACKNOWLEDGEMENTS

This project has received funding from the EU's H2020 research and innovation programme under MSCA GA 675324.

REFERENCES

1. Cooke M, King S, Garnier M, Aubanel V. The listening talker: A review of human and algorithmic context-

induced modifications of speech. Comput Speech Lang. 2014;28(2):543–71.
2. Summers W Van, Pisoni DB, Bernacki RH, Pedlow RI, Stokes MA. Effects of noise on speech production: Acoustic and perceptual analyses. J Acoust Soc Am. 1988;84(3):917–28.
3. Junqua J. The Lombard reflex and its role on human listeners and automatic speech recognizers. J Acoust Soc Am. 1993;93(1):510–24.
4. Kent RD. Research on speech motor control and its disorders: A review and prospective. J Commun Disord. 2000;33(5):391–427.
5. Ziegler W. Task-related factors in oral motor control: Speech and oral diadochokinesis in dysarthria and apraxia of speech. Brain Lang. 2002;80:556–75.
6. Duffy, Joseph R. Motor Speech Disorders: Substrates, Differential Diagnosis, and Management. 3rd ed. St. Louis: Elsevier; 2013.
7. Bernthal JE, Bankson NW, Flipsen PJ. Articulation and Phonological Disorders: Speech Sound Disorders in Children. 6th ed. Boston, Mass: Pearson/Allyn & Bacon; 2009.
8. Maas E. Speech and nonspeech: What are we talking about? Int J Speech Lang Pathol. 2017;19(4):345–59.
9. Kryter KD. Methods for the calculation and use of the Articulation Index. J Acoust Soc Am. 1962;34(11):1689–97.
10. Kryter KD. Validation of the Articulation Index. J Acoust Soc Am. 1962;34(11):1698–702.
11. Steeneken HJM, Houtgast T. A physical method for measuring speech-transmission quality. J Acoust Soc Am. 1980;67(1):318–26.
12. Cooke M. A glimpsing model of speech perception in noise. J Acoust Soc Am. 2006;119(3):1562–73.
13. Tang Y, Cooke M. Optimised spectral weightings for noise-dependent speech intelligibility enhancement. Proc Interspeech 12; 9-13 September 2012; Portland, USA 2012. p. 955–8.
14. Valentini-Botinhao C, Maia R, Yamagishi J, King S, Zen H. Cepstral analysis based on the glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise. Proc ICASSP 12; 25-30 March 2012; Kyoto, Japan 2012. p. 3997–4000.
15. Tang Y, Cooke M. Glimpse-based metrics for predicting speech intelligibility in additive noise conditions. Proc Interspeech 16; 8-12 September 2016; San Francisco, USA 2016. p. 2488–92.
16. Boersma P, Weenink D. Praat: doing phonetics by computer [Computer program]. Version 6.0.36. 2017.