COMPARING IDENTIFICATION OF ARTIFICIAL AND NATURAL VOWELS

Sophie ter Schure, Kateřina Chládková & Jan-Willem van Leussen

Amsterdam Center for Language and Communication, University of Amsterdam, the Netherlands
s.m.m.terschure@uva.nl; k.chladkova@uva.nl; j.w.vanleussen@uva.nl

ABSTRACT

In this study, we investigate how listeners classify non-native vowels, comparing classification of synthetically produced stimuli to that of natural stimuli. A forced choice identification task reveals that synthetic vowels are labeled differently from natural vowels and take more processing time. Participants are also less concordant in labeling synthetic tokens as compared to natural tokens. Because listeners' performance on and classification of synthetic and natural vowels is different, speech perception studies with synthetic stimuli should be cautiously interpreted in terms of how humans perceive the sounds of their natural language.

Keywords: vowel perception, forced-choice classification, natural stimuli, synthetic stimuli

1. INTRODUCTION

Since scientists succeeded in creating intelligible synthetic speech in the first half of the last century (see [9] for a review), countless studies have used artificial speech sounds to test hypotheses about natural speech perception (e.g. [1, 5, 7, 8]). A great advantage of using formant synthesis for categorization tasks (rather than unit-based or statistical-parametric synthesis) is that parameters of interest can be systematically varied while irrelevant parameters are kept completely constant. By varying e.g. duration or formant frequency in equidistant steps, the effects of various acoustic cues on stimulus response can be isolated.

Using systematically varied synthetic stimuli, many important aspects of speech perception have been investigated. For instance, it has been shown that, holding acoustic differences constant, discrimination across phoneme categories is easier than within categories [8]; also, that the vowel space seems to be 'warped' so that close to the prototypical instance of a specific vowel, differences between stimuli are more difficult to perceive than differences between stimuli far from the prototype [7]; that discrimination of phonetic continua is language dependent [8], that vowels can be identified solely on the basis of consonant-vowel transitions [14], and that listeners prefer auditorily peripheral speech sounds [5].

Many studies have used natural (or natural manipulated) tokens (e.g. [2, 11]) to investigate speech perception. However, some basic findings of the studies with synthetic stimuli listed above have, to our knowledge, never been replicated with natural speech [5, 8]. Although some research indicates that listeners perform equally well on synthetic and natural vowels [12], other studies suggest that identification of natural vowels is better than that of synthetic ones, even if these are carefully modeled after natural speech [4].

On the one hand, synthetic speech sounds are audibly different from natural speech [13]. On the other hand, precise control over stimulus properties is desirable to investigate the role of phonetic detail in speech perception.

To find out whether there is a difference between listeners' perception of natural and synthetic speech sounds, we presented a multiple forced choice (MFC) identification task containing both natural and artificial vowels to Dutch-speaking participants. Both sets of vowels came from a source unfamiliar to the participants: the natural stimuli were produced by speakers of Czech, while the artificial stimuli were produced through Klatt synthesis [6].

2. PERCEPTION EXPERIMENT

2.1. Participants

Twenty-five native speakers of Dutch (14 females, mean age 22.08, age range 18-28) participated in the study. All were students or recent graduates. To minimize dialectal perception differences, we selected only participants from the western 'Randstad' area of the Netherlands. Furthermore, only participants with limited exposure to foreign languages were selected for the study.

2.2. Stimuli

The *synthetic* stimuli were sampled from the whole range of possible values: F1 ranged from 260 Hz to 1200 Hz and F2 from 800 Hz to 3000 Hz. Both F1 and F2 were sampled in 16 perceptually equivalent steps (on the Erb scale). We excluded 62 tokens: those for which F1 would be equal to or higher than F2, and non-human sounding tokens with both high F1 and high F2. The resulting F1-F2 vowel grid contained 194 tokens. Each of these tokens was synthesized with three different F3 values: 2900 Hz, 3277 Hz, and 3700 Hz.[1]

This procedure yielded a total of 582 synthetic stimuli (Figure 1). Each token had a duration of 148.5 ms. Stimuli were modeled after a female voice (with a rise-fall contour from 220 to 270 to 180 Hz) and Klatt-synthesized in Praat [3].

Figure 1: F1-F2 plane with the 582 synthesized tokens; each point was synthesized with three distinct F3 values.

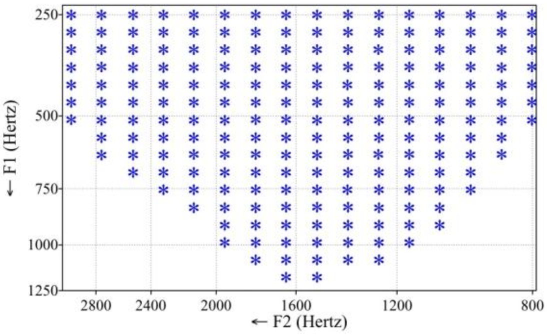
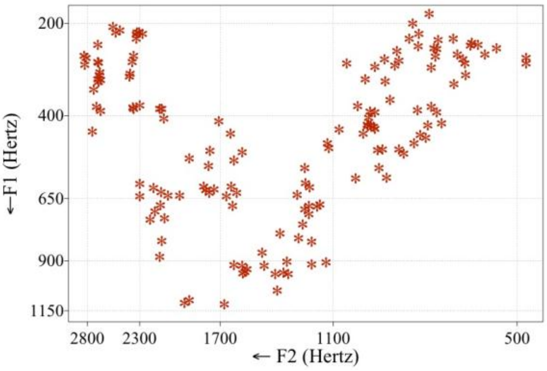


Figure 2: Plot of the 150 natural tokens on the F1-F2 plane.



The *natural* stimuli were extracted from recordings[2] of 10 young monolingual speakers of Czech (5 females). We used isolated vowels from the final position of the phrase *Ve slově CVC máme V* ("In the word CVC we have V"). We selected three tokens of each of the five Czech short vowel categories /a ɛ i o u/ per speaker, choosing those tokens that were closest in duration to the synthetic stimuli (median duration of stimuli in the natural set was 153 ms). A native Czech listener identified all natural stimuli as the intended vowel category. In total, the stimulus set consisted of 150 natural tokens (Figure 2).

2.3. Task

Participants were tested on two MFC identification tasks run in Praat [3], in a soundproof room. Stimuli were played through Sennheiser HD 25 headphones connected to an Edirol UA-25 sound card. We asked subjects to label each stimulus as one of 15 Dutch vowels /i y ɪ ʏ ø e ɛ a ɑ ɔ o u ɛi œy ɔu/ by clicking response buttons on the screen. These contained orthographic representations of the vowels in a bVt or pVk word (e.g. *bot* 'bone', *pauk* 'kettle drum'). A practice task with 15 stimuli preceded the experiment.

In the first task, participants were told that the stimuli were vowels cut from recordings of a Dutch speaker. In fact, they heard the artificial stimuli. This task was interspersed by three breaks. Participants were told that the next task was the same, but that stimuli now came from recordings of different Dutch speakers. This time participants heard the natural Czech stimuli. In both tasks, stimulus order was randomized for each subject.[3]

3. RESULTS & DISCUSSION

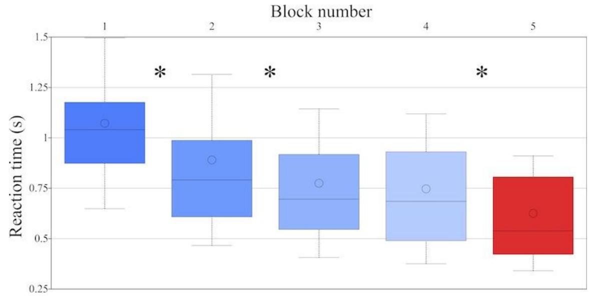
The F1 and F2 ranges across the 11 speakers (10 human, 1 artificial) were not the same. Therefore, before statistically evaluating listeners' performance on the different stimuli, we normalize the vowel space per speaker, using the *z*-score procedure of [10]. Such normalization is warranted because speech perception research has shown that listeners normalize for speaker identity [14].

We then assess the differences in listeners' performance in three ways. First, we measure their response time (RT) in identifying the stimuli; second, we test whether vowel choice is dependent on condition if the variance explained by formant frequency and duration is accounted for; lastly, we measure participants' concordance on vowel choice for the synthetic and natural stimuli.

3.1. Response times

RT is calculated from the offset of the stimulus to the moment of the participant's response. We group RTs into 5 blocks. Blocks are naturally divided by the location of pauses during the experiment and contain 146-150 stimuli each. The first four blocks consist of synthetic stimuli, and the last block of natural stimuli. Figure 3 shows the average RTs for each block.

Figure 3: RTs for the synthetic (boxes 1-4) and natural stimuli (box 5). Boxes span Q1-Q2 and Q2-Q3; circles depict the mean, whiskers 1 SD from the mean. Asterisks mark significant between-block differences.



Per subject, we compute a median RT for each block. We then conduct a repeated measures ANOVA with median RT per block as the dependent variable and with block as the within-subjects factor with 5 levels. The analysis reveals a main effect of block ($F[4,96] = 37.259$, $p < 0.001$). Pairwise comparisons show that RT in block 1 is higher than RT in all other blocks (all p's < 0.001). Similarly, RT in block 2 is higher than RT in each of the blocks 3-5 (for each, $p < 0.01$), and RT in block 4 is higher than in block 5 ($p < 0.001$). No significant difference is detected between blocks 3 and 4 ($p = 0.196$).

These results imply that participants are improving at the start of the experiment, which is likely because they are becoming acquainted with the stimuli and the locations of the labels. After three blocks, minimum RT is reached and no longer decreases for synthetic stimuli. However, listening to natural stimuli instead of synthetic ones does further decrease response time.

A further test comparing the RT differences shows that the RT change from block 1 to 2 is 19% larger than the change from block 3 to 4 ($p = 0.001$), which in turn is 13% smaller than the change from block 4 to 5 ($p = 0.039$). Since the RT difference between blocks 4 and 5 is significantly larger than the RT difference between blocks 3 and 4, the smaller RT in block 5 cannot be attributed solely to the training effect. We conclude that this further improvement is caused by the fact that stimuli in block 5 were natural vowels.

3.2. Regression analysis

To test whether condition has a significant effect on category choice after the variance explained by the varying acoustic dimensions is accounted for, we use a multinomial logistic regression analysis for each of the participants separately. Vowel choice (the 15 response categories) was the dependent variable; independent variables were F1, F2, F3, duration of each stimulus, and condition (natural or synthetic).

Table 1 gives the results of a typical participant, showing that the largest influence on category choice is that of F2, after which F1 contributes most. Condition has the third largest influence on category choice, more than F3. This holds for 18 of the 25 participants. For the other 7, F3 comes before condition, but duration does not add to the variance significantly. As expected, adding duration never yielded a better fitting model, since our stimuli were roughly similar in duration

Table 1: Stepwise (forward entry) regression analysis. Effects that explain a significant part of the variance are entered from largest to smallest contributing effect.

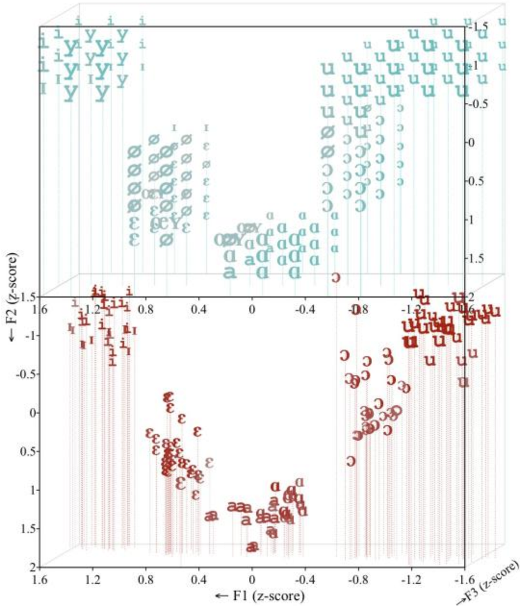
Model	Effects	*-2 Log L.h.*	χ^2	*df*	*p*
0	Intercept	3092.77			
1	F2	1919.72	1173.05	10	<0.001
2	F1	902.12	1017.60	10	<0.001
3	condition	760.83	141.28	10	<0.001
4	F3	710.41	50.42	10	<0.001

3.3. Concordance

Finally, we perform a paired t-test on the amount of agreement on the natural and the synthetic tokens. Each listener labeled each stimulus once, yielding 25 labels per stimulus. For every stimulus, we compute the most-given label, and how often this label was assigned (i.e. the agreement score).[4]

The fact that the natural stimuli did not span the whole vowel space (e.g. no tokens in the mid-central region) might bring a bias into the comparison of synthetic and natural agreement scores (e.g. listeners may be less sure about the nature of a schwa-like vowel than about the nature of a more peripheral vowel). Therefore, in the present comparison, we include only those synthetic stimuli that lie within 2 SDs of the mean values for the 5 intended natural vowels after normalization. This selection process yields 165 synthetic tokens that are, in their location in the speaker-normalized F1-F2 plane, comparable to the natural ones (Figure 4).

Figure 4: Isometric projection of participants' agreement on synthetic (top) and natural stimuli (bottom). Symbols convey the most frequently reported vowel category; darker symbols encode higher agreement. Symbols further on the z-axis have a higher F3.



The difference between the groups is significant: $t[314] = 7.834$, $p < 0.001$. Specifically, natural stimuli yield a 17% higher agreement score than synthetic stimuli (CI = 12.6-21.1).

4. CONCLUSION

Our analysis shows that classifying synthetic stimuli is not the same as classifying natural ones: response time for synthetic vowels is higher, suggesting processing synthetic speech takes longer; categorization is dependent on whether vowels are natural or synthetic, even after formant and duration differences between the stimuli are accounted for; and participants are less congruent about their category choice for synthetic vowels.

The synthetic tokens used in the present comparison did not model all the acoustic properties of our natural tokens, which is why we did not compare the actual response labels given to stimuli in the two sets. Nevertheless, our results show that listeners' responses, as well as their performance in terms of reaction time and congruence, are condition-dependent.

Research with synthetic stimuli has contributed substantially to our current understanding of speech perception; however, our findings suggest that some caution is warranted when generalizing findings obtained with synthetic stimuli to natural speech perception.

5. REFERENCES

[1] Abramson, A.S., Lisker, L. 1970. Discriminability along the voicing continuum: cross-language tests. *Proc. 6th ICPhS* Prague, 569-573.

[2] Best, C.T., McRoberts, G.W., Lafleur, R., Silverisenstadt, J. 1995. Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant Behavior and Development* 350, 339-350.

[3] Boersma, P., Weenink, D. Praat: Doing phonetics by computer. [Computer program], retrieved from *http://www.praat.org*.

[4] Hillenbrand, J.M., Nearey, T.M. 1999. Identification of resynthesized /hVd/ utterances: Effects of formant contour. *J. Acoust. Soc. Am.* 105, 3509-3523.

[5] Johnson, K., Flemming, E., Wright, R. 1993. The hyperspace effect: Phonetic targets are hyperarticulated. *Language* 69, 505-528.

[6] Klatt, D.H., Klatt, L.C. 1990. Analysis, synthesis and perception of voice quality variations among male and female talkers. *J. Acoust. Soc. Am.* 87, 820-856.

[7] Kuhl, P.K. 1991. Human adults and human infants show a "perceptual magnetic effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50, 93-107.

[8] Ladefoged, P., Broadbent, D.E. 1956. Information conveyed by vowels. *J. Acoust. Soc. Am.* 29, 98-104.

[9] Linggard, R. 1985. *Electronic Synthesis of Speech*. Cambridge University Press.

[10] Lobanov, B.M. 1970. Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am.* 49, 606-608.

[11] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264, 746-748.

[12] Morton, J., Carpenter, A. 1962. Judgement of the vowel colour of natural and artificial sounds. *Lang. Speech.* 5, 190-205.

[13] Nusbaum, H.C., Francis, A.L., Henly, A.S. 1995. Measuring the naturalness of synthetic speech. *International Journal of Speech Technology* 1, 7-19.

[14] Rakerd, B., Verbrugge, R.R. 1987. Evidence for talker-independent information for vowels. *Lang. Speech* 29, 39-57.

[1] F3 was always at least 200 Hz above the token's F2. Higher formants were added in a similar fashion to create a flatter spectrum.

[2] The recordings were made in a sound-treated booth with a Røde Broadcaster microphone (cardioid), a Mackie 1642-VLZ3 mixer, and an M-audio Delta 66 computer sound card (44.1 kHz sampling rate and 32 bits quantization).

[3] The natural task always followed the synthetic task, as we felt that the change from natural to synthetic would make the synthetic nature of the stimuli more obvious than if synthetic were presented first, while we wanted to keep the participants ignorant of the stimulus type.

[4] For instance, if 10 listeners labeled a particular stimulus as /i/, 9 listeners as /ɪ/, and 6 listeners as /e/, then the agreement score for that stimulus was 40%; if 23 listeners labeled a stimulus as /i/ and 2 listeners as /ɪ/, the agreement score was 92%.