Using systematic synthetic voice differences to probe the perception of mixed-talker vowels

Santiago Barreda & Terrance M. Nearey

Alberta Phonetics Laboratory

Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada | Canada

Introduction

It is well known that vowel identification performance is worse on mixed-speaker lists relative to blocked-speaker lists. However, previous experiments have found mixed-talker lists made up of 'dissimilar' voices with different formant spaces were not as difficult as those made up of 'similar' voices with different formant spaces [1, 2].

This might be expected if speaker normalization were an active process where a listener has to 'decide' whether to normalize each vowel in a round. However, this would not be expected if vowel normalization were an automatic process [3,4].

Objective: To investigate the effects of different context voices on the ability to correctly identify the vowels of a base voice by systematically varying the **formant spaces (FS)** and f0+source of the voices in a round. **The experiment outlined here is an extension of [1,2].**

Methods

Participants: 71 native English speakers. Each participant was randomly assigned to each one of four target vowel groups.

Stimuli: The stimuli were made up of the vowels / **æ ɑ ʊ ʌ** / from 6 different 'voices', each of which differed in terms of their FS and/or f0+source characteristics. **The FS shifts** consisted of a 10% or 20% **increase of all formant frequencies** relative to baseline. The two **f0+source** levels were an f0 of 120 Hz with modal source characteristics and an f0 of 240 Hz with breathy source characteristics.

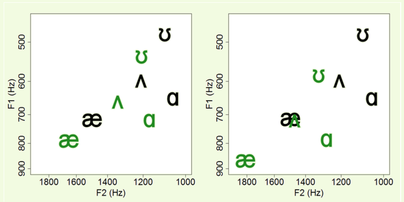

Figure 1 *– In the left panel, the vowels of the baseline FS (black), are compared to those of the 10% shifted FS (green). In the right panel, the baseline FS (black) is compared to the 20% shifted FS (green).*

Procedure: Participants performed a speeded monitoring task where they responded when they heard the target vowel and to ignored any other vowel. The experiment was split into 42 rounds where each round was made up of two voices. All combinations of the 6 voices were used, with repetition, twice each. Each round consisted of 30 vowels (12 targets + 18 distractors), played with an 800 ms inter stimulus interval.

Analysis: The performance for each voice in every round was considered independently. In all cases, the 'base' voice refers to the voice whose performance is being discussed while the 'context' voice refers to the other voice in the round. The results of the / ʌ / target group were not used due to a high level of inconsistency in the responses. The responses of 6 participants were also removed due to poor performance leaving 52 participants (18 in the / ʊ / group and 17 in the / æ / and / ɑ / groups).

Results

The last 15 participants performed an additional task. At the end of each round, they were asked if the round contained **one voice** or **more than one voice**, and whether they were **confident** or unsure of this. A summary of the results is presented in Figure 2.

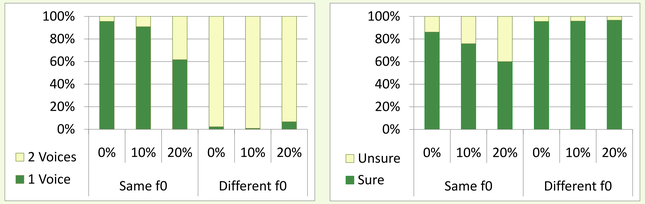

Figure 2 *– Percent of rounds in which participants indicated hearing two voices and whether they were sure of this. Percentages below columns refer to the shift in the formant spaces between the voices that make up the round.*

- FS shifts alone were not enough to give the impression of more than one talker, even with a 20% shift participants heard two voices in only 39% of cases.
- An **f0 difference** in the two voices in a round resulted in the impression that there were **two voices**, *even when they had the same formant space*.
- An f0 difference also resulted in confidence that there were two voices, while formant space differences alone resulted in doubt as to the number of speakers.

Response times. The effect of the context voice on the response times for the base voice with the lowest FS and f0 (**Male voice**) and for the base voice with the highest FS and f0 (**Female voice**) was modelled. Reaction times reported are for correct identifications only.

A linear mixed-effects model was fit to the data where the f0 of the context voice and the FS shift of the context voice were the fixed effects and the participant was the random effect. This was done independently for each target vowel group. Results are presented in Table 1.

	Male Voice	
	f0 Difference	FS Shift
æ	**30.5 (3)**	**18.8 (3)**
ɑ	**32 (4.3)**	7.2 (1.6)
ʊ	**30.6 (4)**	**9.4 (2)**

	Female Voice	
	f0 Difference	FS Shift
æ	**29.5 (4.1)**	.5 (.11)
ɑ	**28.9 (3.3)**	4.2 (.91)
ʊ	**24.6 (3.1)**	**14.3 (3)**

Table 1 *– Estimated effects of context voice f0 differences and FS shifts on reaction times for different base voices, in milliseconds. The numbers in brackets are the t-values associated with each effect. f0 was coded as 0 = same, 1 = different and FS shifts were coded as 1 = same, 2 = 10% shift, 3 = 20% shift..*

- In all cases, **differences in the context voice increase response times for the base voice**.
- Although FS shifts only have a significant effect on reaction times half of the time, **an f0 difference always has a significant effect on response times**.
- **There is remarkable stability in the effect of f0 on response times**. The estimate of this effect varies by only 6 ms across all target vowel groups and for both the male and female voices.

Identification accuracy. The effect of the context voice on the identification of vowels from base voices was also considered for the Male and Female voices. Identification performance was measured using d-prime, which takes into account both correct identifications and false alarms.

A linear mixed-effects model was fit to the data where the f0 of the context voice and the FS shift of the context voice were the fixed effects and participant was the random effect. The coding used for this model is the same as that given in Table 1. Results are presented in Table 2 and visually in Figure 3.

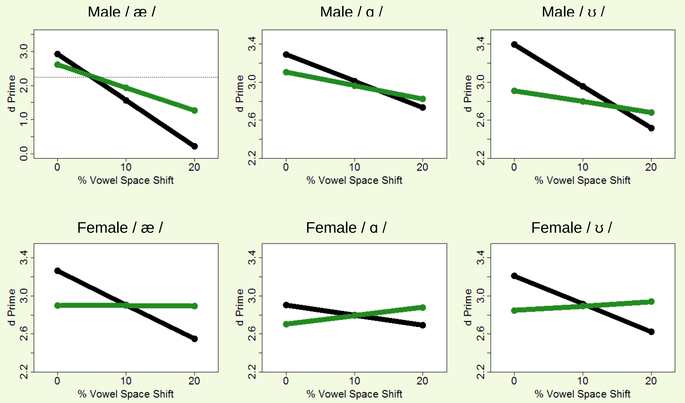

Figure 3 *– Performance for the male and female voices as a function of the properties of the context voice, presented by target vowel. Line colours indicate whether the context voice had the same (black) or different (green) f0+source properties. All panels have the same range except / æ / for the male voice. The range for this vowel is much larger; the dotted line in this panel indicates the limit on the range of the other panels.*

	Male Voice		
	f0 Difference	FS Shift	f0*FS
æ	**-.99 (-3.7)**	**-1.3 (-7)**	**.68 (5.3)**
ɑ	-.32 (-1.5)	-.27 (1.85)	.13 (1.36)
ʊ	**-.80 (-3.4)**	**-.43 (-2.5)**	**.32 (2.7)**

	Female Voice		
	f0 Difference	FS Shift	f0*FS
æ	**-.71 (2.9)**	**-.35 (-2)**	**.35 (3.3)**
ɑ	-.4 (1.57)	-.1 (-.57)	.19 (1.76)
ʊ	**-.7 (-2.8)**	-.3 (-1.6)	**.33 (3)**

Table 2 *– Estimated effects of context voice f0 differences, FS shifts and the interaction of the two (f0*FS) on performance as measured by d-prime. Numbers in brackets are t-values associated with the estimates. Effects were coded in the manner outlined in the caption for Table 1.*

- Unlike for response times, there were significant interaction between f0 and FS shifts, vowel target and performance.
- Overall, **An f0 difference hurts performance when two voices have identical FS but improves performance when FS are very different**.
- Performance was most varied when targeting the male / æ /; this was the vowel with the closest F1-F2 values to another shifted vowel.
- Performance when targeting / ɑ / was least affected by the context voice. This vowel also showed the least variation in performance overall. It is also the vowel with the least overlap with the other target and distractor vowels.

Conclusion

The goal of this experiment was to extend the findings reported in [1,2]. The results presented here confirm those findings. We used targets in the interior of the vowel space, rather than on the periphery. As a result we got much lower identification rates and greater variability in performance.

- **Speaker differences can have a non-additive effect on identification performance** where, in some cases, larger differences between voices facilitate identification.
- **This effect is largely a result of f0**. Listeners appear to rely on this to separate voices.
- **Even large FS differences do not appear to be enough** to help listeners pick voices apart.
- There were **significant target x FS shift x f0 interactions** for performance, but no such interactions for response times.
- The effects on **identification** performance reported are **dependent on the specific locations of the targets used within the FS**.
- Reaction times are less dependent on the specific vowels used.
- All **effects** reported here are **purely extrinsic;** the effects of a context voice on the identification of a base voice.

References

[1] Nusbaum, H. C., & T. M. Morin. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech perception, speech production, and linguistic structure*. Tokyo: OHM. 113-124.

[2] Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance* 33: 391-409.

[3] Smith, David R. R. & Roy D. Patterson. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. Journal of the Acoustical Society of America 118: 3177-3186.

[4] Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H. & Toshio Irino. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America* 117: 305-318.

Contact Information

Santiago Barreda (sbarreda@ualberta.ca)

Terrance M. Nearey (t.nearey@ualberta.ca)

Department of Linguistics, 4-32 Assiniboia Hall, University of Alberta, Edmonton, Alberta, Canada T6G 2E7