MODELING THE PERCEPTION OF VOWEL NASALIZATION

Yongqing Ye

Michigan State University
yeyongqi@msu.edu
Website: https://yongqingye.github.io

Acknowledgement

I'd like to thank my advisor, Dr. Karthik Durvasula, for his guidance, the experimental participants for their contributions, and the MSU PhonoGroup for their feedback and support, all of which made this project possible.

Bayesian Perception

Perception Models:

- Bayesian models that employ step-conditioned distributions for two categories of vowels -- phonologically nasalized vowels and oral vowels.
- Bayes' Theorem

$$P(A\,\big|\,B) = \frac{P(B\,|\,A) \times P(A)}{P(B)}$$

$$P(Category_i\,\big|\,Acoustics) = \frac{P(Acoustics\,|\,Category_i) \times P(Category_i)}{\sum_{i=1}^{n} P(Category_i) \times P(Acoustics\,|\,Category_i)}$$

Acoustic models:

- Acoustic models were trained for two vowel categories: oral (V-oral) and nasalized (V-nasalized).
- Mel-frequency cepstral coefficients (MFCCs) were used to represent acoustic information (Davis and Mermelstein, 1980).
- Likelihood distributions were generated for each category.
- The models were tested with the same nonce words as presented to actual listeners.

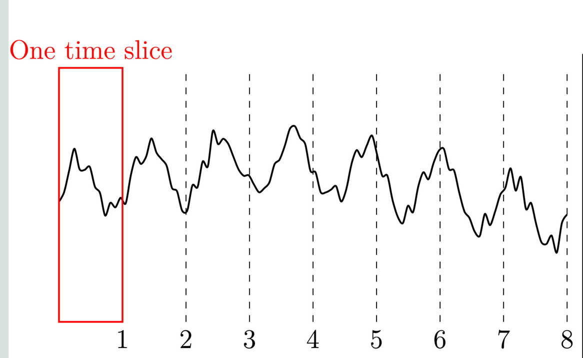


Fig. 1 Dividing a vowel into eight discrete time slices and measuring 12 MFCCs from each time slice

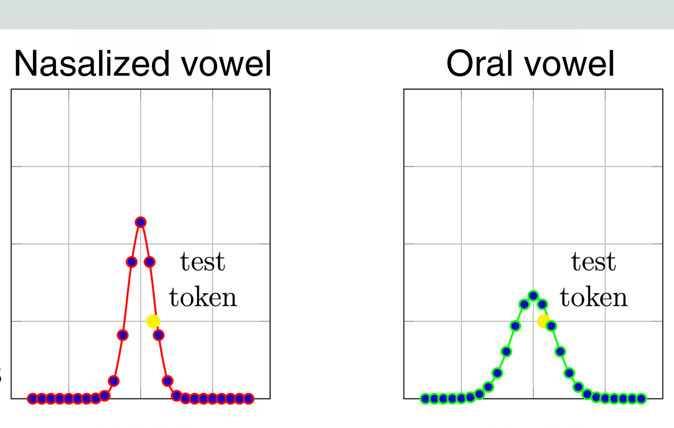


Fig. 2 Simulated probability distributions for two vowel categories at each time point for a single MFCC value

Introduction

Some debates around speech perception:

- Whether listeners access stored exemplars with fine-grained phonetic details (Pierrehumbert et al., 2002; Pierrehumbert, 2001, 2016). or are attuned to subtle variations such as different levels of coarticulation (Fowler, 1981, 1984).
- Whether listeners perceive speech through underspecified URs or SRs (Marslen-Wilson 1980, 1984, 1991; Ohala and Ohala, 1995; Kotzor et al. 2022).
- What is the mechanism of speech perception (Liberman et al. 1967; McClelland and Elman, 1986; Cutler and Norris 1979; McClelland and Elman 1986; Norris and McQueen 2008 etc.).

Questions

Do listeners rely on ...

- Incremental inference updates or immediate acoustics?
- Nuanced temporal information or abstract categories?
- Underspecified representations?

Take-aways

1. **Categorical, underspecified representations are sufficient to account for the patterns in listeners' perception of vowel nasalization.**
2. **Listeners engage in Bayesian inference where their decisions are continuously updated based on previous information.**

Methodology

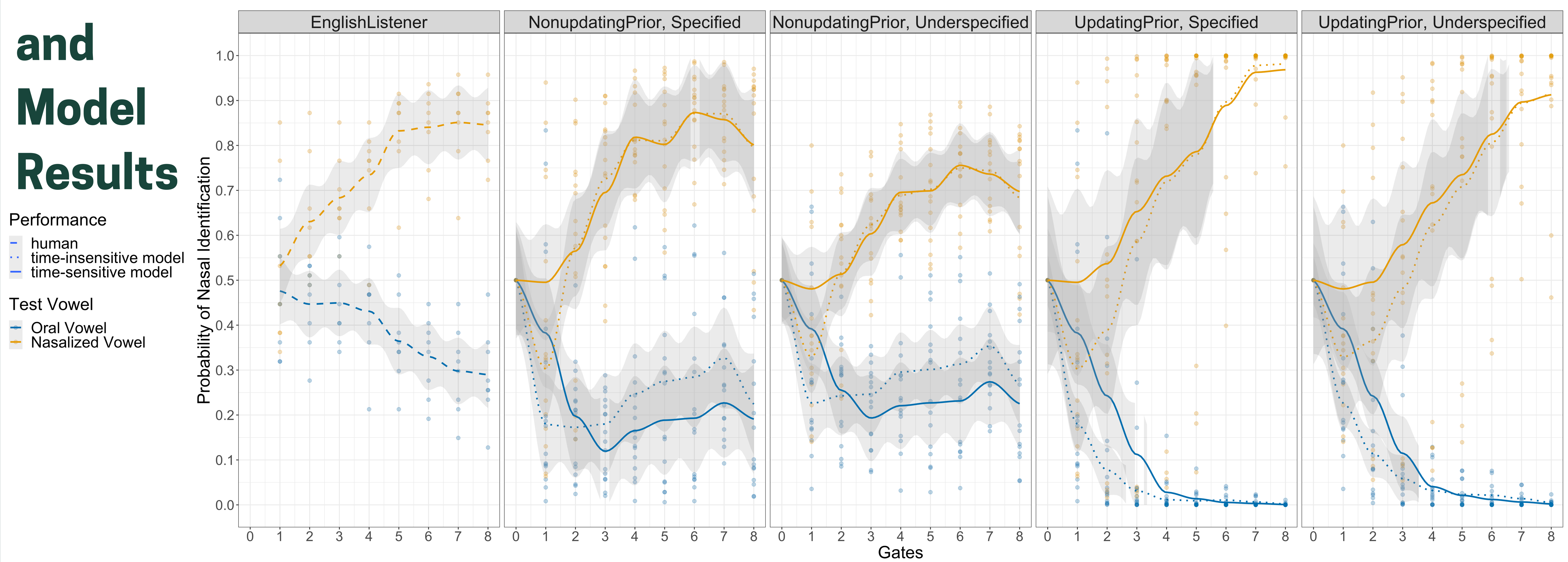
- Participants: 43 native American English speakers.
- Production Task:
  - Wordlist of CVC (oral) and CVN (nasalized) words + fillers.
  - Multiple repetitions of each word.
  - Data used to train acoustic models.
- Perception Task:
  - Forced-choice segment identification task on end-truncated CVC and CVN nonce words.
  - Audio stimuli presented in 8 gated time steps (gates).

Listener and Model Results

Fig. 3a Percentage of nasal responses from 43 English speakers

Fig. 3b-c Bayesian Perception Model with Nonupdating Prior

Fig. 3d-e Bayesian Perception Model with Updating Prior



Listener and Model:
(comparing 3a and 3b-e)

- Listeners could differentiate between V-oral and V-nasalized from the outset.
- Listener accuracy is monotonically increasing.
- Some models perform better than listeners.

Time Sensitivity:
(comparing linetypes)

- **Yes:** Each vowel category is represented with eight time-normalized multidimensional distributions.
- **No:** One distribution across the entire vowel.

Time-insensitive models perform well.
→ Listeners do not need fine temporal details for perceptual accuracy.

Underspecification:
(comparing 3b and 3c; 3d and 3e)

- **Yes:** V-oral was trained with general data from all words (CVC+CVN), V-nasalized was trained with CVN words only.
- **No:** V-oral and V-nasalized were trained with context-specific data from CVC and CVN words respectively.

Models using underspecified representations match the performance from specified acoustic models.
→ Listeners do not need fully specified representations.

Updating Inference:
(comparing 3b-c and 3d-e)

- **Yes:** The posterior at each time step/gate becomes the prior for the next time step/gate.
- **No:** Unchanging, equiprobable priors.

Models with dynamically updating priors, better reflected the monotonic increasing trend in the actual listener behavior than the non-updating ones.
→ Listeners' decisions are continuously influenced by an evolving prior that incorporates newly accumulated probabilistic information .

What's Next?

- Modeling perception in languages with different vowel nasalization patterns: nasal vowels (e.g. Hindi); phonetic nasalization (e.g. Peninsular Spanish).
- Quantitatively compare model performance.
- Evaluating the assumptions and predictions of perception theories.

Selected References

Feldman, Jacob (2009). "Bayes and the simplicity principle in perception." In: Psychological review 116.4, p. 875.
Fowler, Carol A (1984). "Segmentation of coarticulated speech in perception". In: Perception & Psychophysics 36.4, pp. 359–368.
Norris, Dennis and James M McQueen (2008). "Shortlist B: a Bayesian model of continuous speech recognition." In: Psychological review 115.2, p. 357.
Pierrehumbert, Janet B (2001). "Exemplar dynamics: Word frequency, lenition and contrast". In:Typological studies in language 45, pp. 137–158.