Development of first language cue weights from error-driven learning of the continuous speech signal

Jessie S. Nixon, Fabian Tomaschek
University of Tübingen, jessie.nixon@uni-tuebingen.de, fabian.tomaschek@uni-tuebingen.de

Infants begin honing perception to the surrounding language in the first few months of life. One proposal is that this occurs through word discrimination (Werker & Tees, 1984). Others argue that perceptual learning begins too early to result from lexical contrasts, particularly minimal pairs (Maye et al., 2002) and that infants learn in an unsupervised way, through statistical/distributional learning. Statistical learning models propose that listeners learn about their language(s) by tracking the frequency of occurrence of various linguistic items. For example, listeners determine whether a cue dimension is important depending on whether the cue dimension is distributed in one cluster or two. Laboratory studies have shown that exposure to different distributions affects categorisation behaviour (Maye et al., 2002). However, several computational models have demonstrated that an unsupervised, purely statistical approach may not be sufficient to model acquisition of speech sounds (Feldman, Griffiths, Goldwater, and Morgan, 2013; McMurray, Aslin & Toscano, 2009; McMurray & Hollich, 2009). Error-driven learning models (e.g. the Rescorla & Wagner, 1972) propose instead that learning occurs through prediction and prediction error. Perceptual cues are used to predict important events. Based on feedback from prediction error, expectations about future events are adjusted. Previous research suggests that second language speech sound acquisition involves cue competition, a key assumption of error-driven models (Nixon, 2020). The present study proposes an error-driven learning model of first language speech sound acquisition.

We investigate whether early infant acquisition of speech cues could occur through error-driven, discriminative learning of the acoustic speech signal. We use a simple two-layer (cue-outcome) Rescorla-Wagner network (Rescorla & Wagner, 1972) trained on a corpus of spontaneous speech in German. Because the model focuses on the first few months of life, no lexical items or *a priori* sound units, such as phonemes or phonetic features, are used as either inputs or outputs of the model. Instead, discretised 25 ms by 0.47 mel spectral components of running speech are used as both input cues and outcomes. Cue weights develop from the cues' informativity for predicting upcoming signal. Learning not only enhances discriminative cues, but also unlearns/ downweights non-discriminative cues. The model output is a matrix of cue-outcome connection weights.

Two tests (AX and AXB) were used to gauge model performance against human behaviour in speech perception tests in the literature, such as infant head-turn decisions. A series of consonant and vowel continua were created. To determine the extent to which each step along the continuum activated the endpoint consonants or vowels (AX test), activation was calculated from the cues across the spectral frequency range. For each sound pair, it was determined whether the left or right endpoint stimulus was more highly activated in each frequency band. The total number of winning frequency bands was then summed for each endpoint to give the probability of response (AXB test). Generalised additive mixed models (GAMMs) showed that sound pairs were discriminated with high accuracy: activation was high to targets and low to competitors. Interestingly, discrimination occurred in the *expected spectral frequency* ranges for the different sound pairs. Yet discrimination does not appear to result from purely acoustic differences, but rather from cue weighting accrued through learning the informative cues in the acoustic signal. For example, activation showed a gradient, but non-linear decrease with increasing acoustic distance from the target, mimicking the categorisation function typically found in AXB tasks (e.g. Best, McRoberts & Sithole, 1988).

The present study proposes an error-driven model of early infant speech sound acquisition. The model assumes no a priori linguistic units. The key assumption of the model is that infants use incoming acoustic signal to predict and discriminate upcoming acoustic signal. The model takes spectral components from natural running speech as both inputs and outputs. After training, simulations of infant head-turn decisions accurately discriminated sound pairs and showed human-like categorisation. In summary, the results suggest that error-driven learning of the acoustic signal may be a feasible alternative to statistical clustering models for infant speech sound acquisition.

References

Best, C. T., McRoberts, G. W., Sithole, N. M., 1988. Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. Journal of experimental psychology: human perception and performance 14 (3), 345

Feldman, N. H., Griffiths, T. L., Goldwater, S., Morgan, J. L., 2013. A role for the developing lexicon in phonetic category acquisition. Psychological review 120 (4), 751.

Maye, J., Werker, J. F., Gerken, L., 2002. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition 82 (3).

McMurray, B., Aslin, R. N., Toscano, J. C., 2009. Statistical learning of phonetic categories:
insights from a computational approach. Developmental Science 12 (3), 369–378.

McMurray, B., & Hollich, G. (2009). Core computational principles of language acquisition: can statistical learning do the job? Introduction to special section. Developmental Science.

Nixon, J. S. (2020). Of mice and men: speech acquisition as discriminative learning from prediction error, not just statistical tracking. Cognition, 197, 104081.

Rescorla, R. and Wagner, A., (1972). A theory of Pavlovian conditioning. Black, A. H., Prokasy, W. F. (Eds.), Classical conditioning II: Current research theory. Ap.-Cent-Crofts, New-York,. 64– 99.