Interspeech 2018
2-6 September 2018, Hyderabad

Effective acoustic cue learning is not just statistical, it is discriminative

Jessie S. Nixon[1]

[1]Quantitative Linguistics, Eberhard Karls Universität Tübingen, Germany

jessie.nixon@uni-tuebingen.de

Abstract

A growing statistical learning literature suggests that listeners extract statistical information from the linguistic environment. However, distributional frequency may be insufficient for important but relatively low-frequency cues. Acquisition of linguistic knowledge may rely not merely on co-occurrences but on predictive relationships between cues and their outcomes. The present study investigates effects of predictive temporal cue structure on acquisition of a non-native acoustic cue dimension.

During training, native English speakers saw coloured shape objects and heard spoken Min Chinese words with six different lexical tones. Tones were the only reliable cue to identifying the associated object. Words also contained a salient cue that did not discriminate between objects. Three tones occurred with high-frequency and three with low-frequency in training. The critical manipulation was the presentation order: either words, containing complex cue structure, preceded object outcomes (discriminative order) or objects preceded words (non-discriminative order).

Generalised linear mixed models showed accuracy was significantly higher in the discriminative order than the non-discriminative order. These results demonstrate that predictive cue structure can facilitate acquisition of a non-native cue dimension. Feedback from prediction error drives learners to ignore salient non-discriminative cues and effectively learn to use the target cue dimension.

Index Terms: discriminative learning, prediction, error-driven learning, learning theory, speech acquisition, lexical tone, Southern Min Chinese

1. Introduction

Given the substantial variability in the organisation of acoustic cues across the world's languages, how are speech cues acquired? As adults we have become native speakers of whichever language(s) were in our environment as infants and children. First language learners eventually learn to accurately discriminate acoustic information to a very fine level of detail when those acoustic cues are discriminative in their native language. Yet when we begin learning a new language in adulthood, we do not usually obtain the same level as native speakers.

Over the past two decades, substantial evidence has accumulated that listeners are highly sensitive to the statistical distribution of acoustic cues in the speech signal. Experiments have shown that categorisation behaviour as well as online processing measures are affected by the number of Gaussian peaks (unimodal vs. bimodal), the distance between peaks and the statistical variance of input distributions [1, 2, 3, 4, 5, 6]. However, there is also substantial evidence that listeners' knowledge does not completely correspond with the information available in the distributional statistics of the language.

This year is the 30th anniversary of Rescorla's (1998) review [7] of Pavlovian conditioning. The review was essentially a plea to the Psychological community to read, rethink and revise their assumptions about learning, along with a summary of some of the basic principles of associative learning. While a number of studies have investigated language acquisition and processing [8, 9, 10, 11] from a learning theory perspective [12], the field of linguistics in general and speech comprehension and acquisition in particular do not on the whole seem to have fully incorporated the insights from learning theory.

One of the aspects emphasised in [7] was the role of prediction in learning. Decades of research on animal learning demonstrate that learning does not simply derive from co-occurence of events. Rescorla considered conditioning to be a process of learning by exposure to the *relations between events* in the environment. Because this learning was the primary means by which organisms learned how to represent the world, this meant that conditioning was necessarily rich and complex, both in terms of the relations represented and in terms of its effects on behaviour - a far cry from a simple reflex response as Pavlovian conditioning was often characterised to be [7]. Importantly, theories of how organisms encode the relations between events in the world emphasise a necessary discrepancy between the actual state of the world and the representation of that state [13, 14, 15, 16, 12, 17]. This last point is often overlooked or downplayed in discussions of statistical learning.

The way that current perceptual information is perceived depends on previous experience. Kamin [18] demonstrated that if an already-learned cue does the job of discriminating between important outcomes, then an additional cue that provides the same information is not learned. That is, despite consistent co-occurrence between the second cue and its outcome, learning of the cue is 'blocked' by the previously learned cue, because there is no uncertainty left to drive learning [7]. Thus, learning can be thought of not simply as an association between co-occurring events, but as a *competitive* process to optimise discrimination in future events. In a sense, cues compete with each other for the job of reducing uncertainty in the environment. When there is no longer uncertainty, there is no longer any opportunity for learning. Rescorla [7] argues that explaining conditioning in these terms, rather than as a reflex, has consequences for all three of what he considers to be the primary issues in the study of learning: the *circumstances* that produce Pavlovian conditioning, the *content* of the learning and the *effects* on behaviour.

This raises the question of what circumstances, if any, can facilitate learning of previously blocked cues. This is the question an adult learner faces when they begin to learn a new language. The learner needs to reverse the process of blocking that has occurred previously when they learned to ignore cues that were uninformative in the native language.

In a recent study, Ramscar and colleagues [11] investigated the role of the predictive temporal structure of cues and outcomes in the acquisition of object categories. The study showed that when learning labels for visual categories ('species of alien'), participants' ability to correctly identify the labels for

10.21437/Interspeech.2018-1024

each category depended not on the statistical structure, which was the same for all groups, but on the predictive structure of the learning events. In particular, it depended on the temporal order of *cues* and *outcomes*. A salient visual cue (the body) corresponded with a particular label 75% of the time, but with a *different* label the other 25% of the time. So, participants had to learn not to rely on this cue. In order to select the correct label, participants had to learn a complex set of more subtle cues. The critical manipulation was the order in which participants saw the stimuli within the trial. Either the category labels ('This is a wug') preceded the complex visual cues of the objects or, vice versa, the objects preceded the labels ('That was a wug').

Results showed that participants in the two conditions did equally well with the high frequency items. However, participants in the discriminative order were significantly more accurate at identifying the low-frequency items. When the visual cues were presented first (discriminative order), participants could generate predictions about the label to follow. If the expected label did not appear, they could adjust their expections (i.e. association weights between cue and outcome) for the following trials. Thus, cues competed for relevance: the non-discriminative cue (object body) was downweighted and the set of discriminative cues were strengthened. However, if the label preceded the object (non-discriminative order), there was no opportunity for cue-competition, resulting in conditional probability learning. Responses were based largely on the salient non-discriminative feature (the object body) that occurred most often with the outcome.

The above results demonstrate that, at least in acquisition of visual semantic categories, learning depends on the predictive structure of learning events. But words can serve as both cues and outcomes of an event. In [11], the label served as the outcome and was considered a non-divisable, featureless chunk and the features of the objects were visually complex and served as cues. However, like the features that make up object categories, spoken language also contains an incredibly rich set of acoustic cues. The cue complexity of spoken language may allow for cue competition and error-driven learning in the same way as has been shown for visual cue competition. The present study investigated whether the temporal predictive structure of learning events also affects acquisition of a non-native speech cue.

1.1. The present study

The present study investigated the effects of predictive order on learning of a non-native acoustic cue dimension, namely English speakers' acquisition of lexical tone. Lexical tone is often a challenge to native English beginner learners of tone languages, since pitch is not used to discriminate between lexical items in English. Cues were spoken Southern Min Chinese words and outcomes were coloured shapes. During training, participants either heard the spoken word cues first, followed by the visual object outcomes (*discriminative order*) or they saw the visual outcomes followed by the cues (*non-discriminative order*). There were high- (75%) and low-frequency (25%) items. In the test, participants had to select which of three spoken words matched a visual object. Given the results of [11], it was expected that low-frequency items would be learned better in the discriminative order. Specifically, it was expected that in the non-discriminative order, learning performance would rely on co-occurrence statistics.

2. Method

2.1. Participants

Participants were 196 native English speakers in the US recruited online via Amazon Mechanical Turk. The experiment took approximately 20-30 minutes and participants were paid a small sum for their participation.

2.2. Stimuli

Visual stimuli (*outcomes*) were images of three coloured shapes (a red circle, a yellow triangle and a blue square). Auditory stimuli (*cues*) were single-character words produced by a native speaker of Taiwan Southern Min Chinese. Stimuli consisted of three different base syllables ('tshe', 'o' and 'phe') and six different lexical tones. Each base syllable (e.g. 'phe') was produced with two different tones (e.g. 'phe_rising' and 'phe_falling'), resulting in six different tonal syllables. Two different tokens of each syllable were used, to create some acoustic variability in the stimuli.

2.3. Experiment design

Three of the tones occurred with high frequency (75% of training trials) and the other three with low-frequency (25% of training trials). Importantly, the base syllable did **not** always correctly predict the target image. The base syllable (e.g. 'phe') that corresponded to a given image in the high frequency stimuli (e.g. 'phe_rising'; red circle) corresponded to a different image in the low frequency stimuli (e.g. 'phe_falling'; blue square). This meant that in order to correctly identify the low-frequency stimuli, participants needed to ignore the more salient cue, the base syllable, and select the image based on its tone.

2.4. Procedure

The experiment consisted of a training phase and a test phase. During the training phase, participants either heard a spoken word followed by an image on screen (discriminative condition) or saw an image on screen followed by a spoken word (non-discriminative condition). Participants simply clicked on the image to continue to the next trial.

In the test phase, on each trial, one image appeared on the screen with three buttons beneath. In the test (unlike the training phase), the number of high- and low-frequency items was equal. Three auditory stimuli were played in random sequence. The stimuli were either the three high-frequency or the three low-frequency stimuli. Each of the three corresponding buttons was highlighted in sequence as each auditory stimulus played. The task was to click on a button to select the auditory stimulus that corresponded to the image.

Participants completed a brief questionnaire about age, gender and language background prior to the experiment and experience of the game afterwards.

3. Analysis and results

The proportion of clicks on the correct target word and the competitor (i.e. the item for which the base syllable corresponded to the image but the tone did not) are shown in Figure 1 for each condition for the high-frequency (left panel) and low-frequency stimuli (right panel). Each panel shows the non-discriminative order (left bars) and the discriminative order (right bars). For the high-frequency stimuli, the correct target word was selected with very high accuracy in both conditions. For the low-

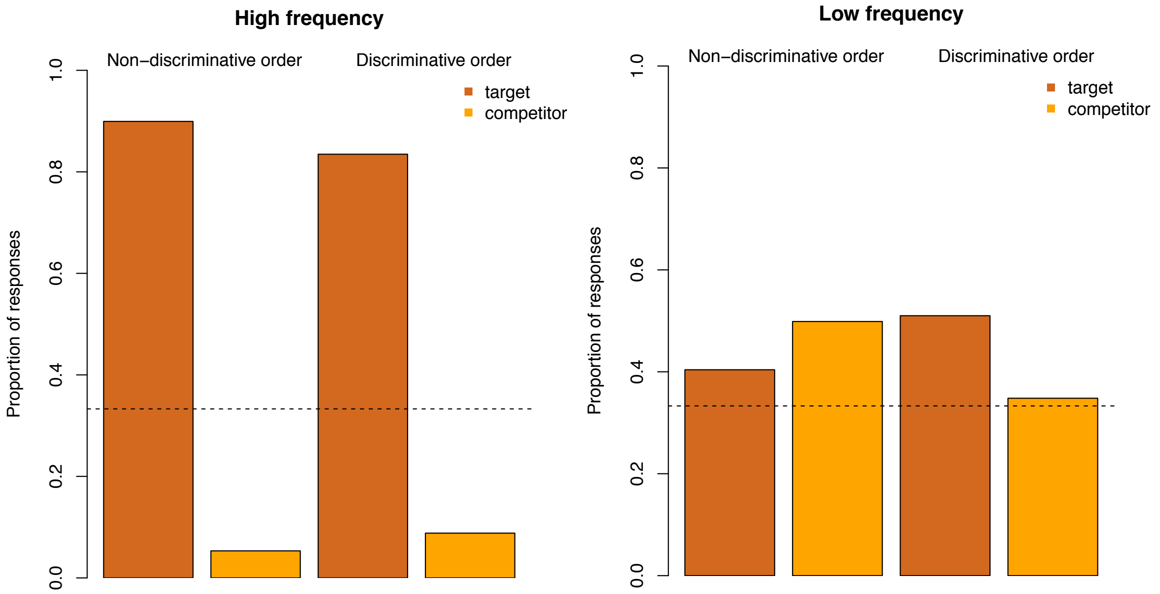

Figure 1: *Proportion of clicks on the target and competitor word buttons per condition for high-frequency (left panel) and low-frequency items (right panel). The horizontal dashed line represents chance level performance.*

frequency stimuli, the target appears to be selected more often in the discriminative order and the competitor appears to be selected more often in the non-discriminative order.

A generalised linear mixed effects (glmer) model was used to test whether the observed differences in accuracy were significant. The model tested the effect of the order of the cue and outcome on the likelihood of participants selecting the target versus competitor items. The model included a two-level factor of condition (discriminative vs. non-discriminative order), a two-level factor of item frequency during training (high- vs. low-frequency) and the two-way interaction, each of which significantly improved model fit. Random intercepts for item and the interaction between participant and frequency were included to account for differences between items, participants and the effect of frequency on participants. Random slopes were not supported. Based on the Ramscar et al. study [11], differences between conditions were expected only in the low-frequency items.

The model summary is shown in Table 1. The dependent variable was the selection decision: competitor (coded as 0) versus target (coded as 1). The discriminative condition for the low-frequency items is on the intercept. There was a significant interaction between condition and frequency. Most importantly, for the low-frequency items, there were significantly fewer target responses in the non-discriminative condition, compared to the discriminative condition. There were more target responses for high-frequency items than low-frequency items, but this did not differ between conditions.

Fixed effects	Estimate	Std. Error	z-value	Pr($<\vert z \vert$)
(Intercept)	0.6596	0.4272	1.544	0.123
cond=ND	-1.1724	0.5336	-2.197	0.028
frequency=high	3.2018	0.6460	4.956	7.19e-07
ND:highFreq	1.8424	0.8104	2.273	0.023

Table 1: *Summary of glmer model. ND = non-discriminative*

4. Discussion

The present study investigated the role of predictive cue structure in learning of non-native speech cues. Participants were presented with visual colour-shape categories and corresponding spoken Southern Min Chinese words. There were two groups of spoken word cues: those that occurred with high-frequency (75%) and those that occurred with low-frequency (25%) during the training. Each base syllable occurred in the high-frequency group with a particular tone and corresponding to a particular shape, and also in the low-frequency group with a different tone and, importantly, corresponding to a *different* shape. Therefore, in order to correctly identify the speech token associated with a particular shape, participants had to ignore the more salient cue of the base syllable and instead make their selection based on the tone.

The critical manipulation was the order of presentation of the cues (spoken words) and outcomes (visual images). The content of the trials was identical between conditions. Results showed that for the low-frequency items, the target spoken word was correctly selected significantly more often when participants were trained with the discriminative order, namely spoken cues before visual outcomes, compared to the non-discriminative order, when the objects preceded the spoken cues. These results demonstrate that in non-native acoustic cue acquisition, simple pairing of objects with their spoken items is not sufficient for effective learning of novel speech cues. Instead, the cue weighting of previously downweighted cue dimensions can be increased with discriminative learning through cue competition.

It is important to note that in both the discriminative and the non-discriminative order, the same information was available. In both cases, a single object was paired with a single word on each trial. In both cases, there was a pause between the two stimulus items, so there was the opportunity to generate a prediction about which stimulus item would follow. In both cases, feedback from prediction error was available when the later stimulus was presented. The difference was that in the discriminative order the candidate cues (e.g. syllable and tone) competed in the process of discriminating between the possible object outcomes. When expectations based on the syllable failed to accurately predict the object outcome, the association between syllable and outcome was downweighted; simultaneously the relative weight of the tone cue increased.

In contrast, in the non-discriminative order, when expectations based on the coloured shape failed to accurately predict the spoken word, the association between the object and word was downweighted; however, critically, there was no relative increase in the weight of other cues. In terms of visual cues, there were no other relevant cues that had potential to increase weight in predicting the word outcomes. In terms of acoustic cues, because the object outcome was already known by the time the word was presented, there was no opportunity to generate predictions based on the acoustic cues, so there was no acoustic cue competition. Therefore, the selection decisions during test were based mainly on the non-discriminative base syllable cue.

The present study of auditory word discrimination replicates the results of a previous study in the visual domain [11]. The results show that associative learning of speech cues does not simply involve pairing of stimuli. This is consistent with Rescorla's characterisation of learning as a process of making predictions about important outcomes and consequently adjusting expectations about how cues predict events [7] and furthermore that this is a discriminative process involving cue competition [11, 19]. In a typical learning event, multiple cues are available in the environment and compete for relevance for predicting important outcomes. When the temporal order of learning events presents cues prior to outcomes, it is possible to generate predictions based on all competing cues and to adjust expectations regarding future events accordingly as a function of prediction error. When a learning event presents the outcome before the cues, there is no opportunity for the cues to compete. This results in learning of co-occurrence statistics [11]. In the present study, this reliance on co-occurrence statistics led participants in the non-discriminative condition to most often select the word containing the salient but non-discriminative cue - the syllable - that most often occurred with the outcome.

In first language (L1) acquisition, acoustic discrimination is emergent. The ability to use bottom-up acoustic cues to predict speech outcomes is less developed when vocabulary size is small; then, as vocabulary size grows and the number of things to be discriminated increases, precision also increases [20]. However, in L2 acquisition, expert knowledge about one's L1 can hinder learning of relevant cues for a new language [21].

In the present study, the cue complexity of the objects and speech stimuli were controlled such that the speech cues were complex and the visual cues were not. However, in the real world, (visual) objects and speech can both occur as both cues and as outcomes. Discrimination of semantic outcomes (e.g. a child learning that a teddy bear is not called a doll) occurs concurrently with discrimination of speech sounds. When an infant sees a bear, predicts the word 'doll' and then subsequently hears the word 'bear', the event facilitates learning to discriminate bears from dolls. When the infant hears the word 'bear' and then their gaze is directed to a teddy bear object, this helps with learning the speech cues that predict this outcome. If the child had expected a *pear*, the error would lower the association weight between particular acoustic cues (e.g. short voice onset time) with the fruit outcome. Both orders will be highly prevalent in a child's input, allowing them to concurrently learn to discriminate things in the world and acoustic speech cues.

In second language acquisition, the semantic knowledge that speakers have already acquired in their first language is often - although not always - useful for acquiring the new language. Having learned to discriminate dolls from bears in one language, the learner expects two labels. Having learned the concept of 'yesterday' in the first language, the learner expects a label for this concept in the second [8]. To the extent that such concepts overlap between the two languages, this expectation increases efficiency of acquiring the words 'bear', 'doll' and 'yesterday' in the new language.[1]

However, in spoken language, expectations based on the sounds of the native language can hinder comprehension and production of speech sounds [21]. Exposure (especially repeated exposure) to cues that are not discriminative in the given learning events will tend to hinder later acquisition of those cues [8]. Acquisition of English involves learning that regardless of the height or direction of the pitch trajectory over a word, the word will have the same meaning. A chair remains a chair, whether it occurs with high, low, rising or falling pitch.[2] English speakers learn to ignore pitch as a lexical cue. The present results suggest that in order to reverse this loss of discrimination, it is advantageous to present the auditory stimulus as cue rather than as outcome.

The present results have interesting consequences for statistical learning studies. In these studies, linguistic input is generally conceptualised as belonging to a static input distribution. Because experience with a cue in a given learning event affects perception of the same cue in subsequent learning events, in order to capture the learning process, learning studies should take into account that the order in which particular stimuli are encountered can have consequences for their discriminative value. Therefore, rather than being sampled from a static distribution of stimuli, perceptual input in the real world evolves through cue competition and discriminative learning.

5. Conclusions

Acquisition of a native language leads to downweighting or loss of discrimination of cues that may be important for learning a second language. The present results show that to reverse this loss of discrimination, listeners can adjust previously learned cue weighting through cue competition. Therefore, in speech acquisition, it is advantageous to present the auditory stimulus as cue rather than as outcome.

6. Acknowledgements

Many thanks to Michael Ramscar, Harald Baayan and four anonymous reviewers for helpful comments on an earlier version of this manuscript. This project was made possible through the support of a subaward under a grant to Northwestern University from the John Templeton Foundation (Award ID 36617) and an ERC Advanced Grant (Grant number 742545). The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

[1] That is not to say that concepts are identical between languages. For example, food is an area where there are large differences in labelling: one language can have a single label for a food type (e.g. 'noodles'), which in another language is differentiated by multiple labels.

[2] In English, changes in pitch at the sentence level (i.e. intonation) act as cues for sentence meaning. But pitch does not discriminate lexical items as it does it tone languages.

7. References

[1] M. Clayards, M. K. Tanenhaus, R. N. Aslin, and R. A. Jacobs, "Perception of speech reflects optimal use of probabilistic speech cues," *Cognition*, vol. 108, no. 3, pp. 804–809, 2008.

[2] P. Escudero, T. Benders, and K. Wanrooij, "Enhanced bimodal distributions facilitate the learning of second language vowels," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, 2011.

[3] J. Maye and L. Gerken, *Learning phonemes without minimal pairs*. Proceedings of the 24th Annual Boston University Conference on Language Development, 2000.

[4] J. S. Nixon, J. van Rij, P. Mok, R. H. Baayen, and Y. Chen, "The temporal dynamics of perceptual uncertainty: eye movement evidence from Cantonese segment and tone perception," *Journal of Memory and Language*, vol. 90, pp. 103–125, 2016.

[5] J. S. Nixon, N. Boll-Avetisyan, T. O. Lentz, S. van Ommen, B. Keij, Ç. Çöltekin, L. Liu, and J. van Rij, "Short-term exposure enhances perception of both between- and within-category acoustic information," in *Proceedings of the 9th International Conference on Speech Prosody*, Poznan, Poland, June 2018.

[6] J. S. Nixon and C. T. Best, "Acoustic cue variability affects eye movement behaviour during non-native speech perception," in *Proceedings of the 9th International Conference on Speech Prosody*, Poznan, Poland, June 2018.

[7] R. A. Rescorla, "Pavlovian conditioning: It's not what you think it is." *American Psychologist*, vol. 43, no. 3, p. 151, 1988.

[8] I. Arnon and M. Ramscar, "Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned," *Cognition*, vol. 122, no. 3, pp. 292–305, 2012.

[9] R. H. Baayen, P. Milin, D. F. urević, P. Hendrix, and M. Marelli, "An amorphous model for morphological processing in visual comprehension based on naive discriminative learning." *Psychological review*, vol. 118, no. 3, p. 438, 2011.

[10] E. Colunga, L. B. Smith, and M. Gasser, "Correlation versus prediction in children's word learning: Cross-linguistic evidence and simulations," *Language and Cognition*, vol. 1, no. 2, pp. 197–217, 2009.

[11] M. Ramscar, D. Yarlett, M. Dye, K. Denny, and K. Thorpe, "The effects of feature-label-order and their implications for symbolic learning," *Cognitive Science*, vol. 34, no. 6, pp. 909–957, 2010.

[12] R. A. Rescorla and A. R. Wagner, "A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," in *Classical conditioning II: Current research and theory*, A. H. Black and W. F. Prokasy, Eds. New-York: Appleton-Century-Crofts, 1972, vol. 2, pp. 64–99.

[13] D. Arnold, F. Tomaschek, K. Sering, F. Lopez, and R. H. Baayen, "Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit," *PloS one*, vol. 12, no. 4, p. e0174623, 2017.

[14] A. Dickinson, *Contemporary animal learning theory*. London, England: Cambridge University Press, 1980, vol. 1.

[15] N. J. Mackintosh, "A theory of attention: variations in the associability of stimuli with reinforcement." *Psychological review*, vol. 82, no. 4, pp. 276–298, 1975.

[16] J. M. Pearce and G. Hall, "A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli." *Psychological review*, vol. 87, no. 6, pp. 532–552, 1980.

[17] K. Sering, P. Milin, and R. H. Baayen, "Language comprehension as a multi-label classification problem," *Statistica Neerlandica*, 2018.

[18] L. J. Kamin, "Attention-like processes in classical conditioning," in *Miami symposium on the prediction of behavior: Aversive stimulation*, 1968, pp. 9–31.

[19] M. Ramscar, M. Dye, and S. M. McCauley, "Error and expectation in language learning: The curious absence of mouses in adult speech," *Language*, vol. 89, no. 4, pp. 760–793, 2013.

[20] V. M. Garlock, A. C. Walley, and J. L. Metsala, "Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults," *Journal of Memory and language*, vol. 45, no. 3, pp. 468–492, 2001.

[21] C. T. Best, "A direct realist view of cross-language speech perception," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.