Tones or Noise:
How Previous Linguistic Experience Influences Lexical Processing of Tone

Amritha Mallikarjun

Abstract

The input people receive in their native language shapes their perceptual understanding of unfamiliar speech. Infants tend to under-generalize their speech input; when presented with words in a specific affect or pitch, they fail to recognize the word if it is subsequently presented in a different affect or pitch (Houston & Jusczyk, 2000, Singh et al., 2008). As infants get older they focus only on the aspects of speech that provide important lexical information to them. For example, Japanese speakers cannot perceive the difference between /r/ and /l/ because their language groups these two sounds together in one phoneme, and as such they have difficulty learning words in English that require an r/l distinction to use properly, like [rip] and [lip]. This study will explore the differences in perceptual understanding between Mandarin speakers and English speakers in a statistical learning paradigm. The participants will listen to words that either have a consistent Mandarin tone associated with them or a random tone. We expect that the Mandarin speakers, who store tone contour as an important aspect of words, would have trouble with an inconsistent tone cue, while English speakers would disregard the tone cue entirely and perform similarly in both conditions.

Introduction

Infants and young children have an extraordinary ability to learn and comprehend language. Regardless of the variation and inconsistencies of the input, they hear the sounds presented to them in fluent speech and manage to segment them into words (Jusczyk & Aslin, 1995). One mechanism that researchers believe is involved in language acquisition is statistical learning, which refers to the process of grouping syllables that co-occur. This sensitivity to probabilistic co-occurrence allows infants to find words in fluent speech based on the probability that certain speech sounds occurred together (Saffran, Aslin & Newport, 1996, Saffran, 2003, Thiessen & Saffran, 2003). For example, an infant may hear the word "pretty" when their parents are talking. If this infant hears this often enough, she will learn that when she hears the syllable "pre-" the syllable "ty" is likely to come afterwards. However, "pretty" can be followed by many different words: pretty baby, pretty eyes, or pretty shoes, as examples. So the infant learns that "-ty" doesn't strongly predict any particular syllable. As such, the infant learns that "pretty" is a word in their language. This mechanism not only works over sounds in speech, but also musical tones, images, and basic syntax structures.

Before infants can successfully perform statistical learning over phonemes, it is necessary for them to perform a more base-level statistical learning on the allophones of the phonemes themselves in order to address variation in their pronunciation and sound. Young infants tend to under-generalize words they hear, storing too much information about their absolute pitch, overall tone contour, rhythmic patterns and timbre, as they are still focusing on the basic acoustic characteristics of language before they learn how this acoustic information is relevant to identifying categories (such as different phonemic categories) in their native language. This under-generalization can be seen in 7.5 month olds that listen to the word "cup" spoken by a female. They then do not recognize "cup" again when produced by a male speaker and treat it as a different word (Houston & Jusczyk, 2000). Singh, Morgan & White (2004) also showed that infants familiarized with words in a happy voice do not listen longer to these words if they are presented in a neutral voice, which would alter the original tone and pitch contour.

While failing to recognize a word when it is presented in a novel pitch or affect would be a maladaptive response for adults, it indicates a representational system that may well be adaptive for language learning. A priori, infants cannot know which aspects of the acoustic signal will be relevant to the native language. For example, while pitch contour is not indicative of lexical meaning in English, it is contrastive in tonal languages such as Mandarin. The fact that infants appear to store "extra" or "irrelevant" information in their early representations of lexical forms may facilitate learning which acoustic features are informative in their native language. Some of this extra information that infants store when they learn words can be useful when they later focus on the characteristics of their native language. For example, infants represent lexical stress from an early age (Jusczyk, Cutler, and Redanz, 1993). Later, they discover that lexical stress is correlated with word boundaries in English. Once they discover this fact, they can use the information they have encoded about lexical stress to help them better identify words (Curtin, Mintz, & Christiansen, 2005).

Eventually, infants learn to generalize across acoustic features that are not relevant to their native language. For example, English-learning infants eventually treat a word presented with two pitch contours as identical, because pitch contour does not signify lexical differences in English. Conversely, an infant learning Mandarin should not generalize across pitch contours, because pitch contour distinguishes between lexical items in their language. This ability for infants to eventually generalize over the extraneous features of sound that do not contribute to their language occurs due to distributional statistics they calculate over the sounds. Even though infants at first have trouble with male and female voice changes as well as affect differences, Singh (2008) has shown that presenting infants with happy, sad, angry, and fearful versions of words helps them to recognize the words when they are presented with happy or neutral affect in an unfamiliar sentence. We can see this phenomenon in the formation of phonemic categories as well. When six-month-olds are presented with either bimodal or unimodal distributional information over a phonetic continuum between /da/ and /ta/, infants only distinguish between these two sounds in the test phase when they heard a bimodal distribution, with frequent peaks at both /da/ and /ta/, rather than only a frequent peak at the blended middle consonant in the unimodal distribution. This again points to the significance of distributional information for the categorization and importantly, the perception of phonemes (Maye, Werker & Gerken, 2002). Given these differing low-level phonemic categorizations, we expect speakers of different languages to also parse different words out of fluent speech via statistical learning based on their previous linguistic input.

Even after the infants learn to generalize over extraneous elements like the gender of the speaker and background sound, they still face a great deal of variability in the phoneme categories depending on their native language. For example, English speakers generally treat the words “sit”, “kitten”, and “tip” as all containing the same phoneme /t/. These sounds, however, are not entirely the same: sit contains the unreleased stop version [t ̚], kitten contains the flap [ɾ], and tip contains the aspirated stop [t^h] (Traeger, 1942). If we swap any of these sounds in these words, however, the meaning remains the same. While it would be strange to hear an over-emphasized aspirated [t^h] in “kitten”, we won’t lose the meaning due to the change in allophone. In contrast, in the national language of Cambodia, Khmer, changing a [t] to a [t^h] does change the meaning of words. While [ta] means “old man”, [t^{h}a] means “say”. This indicates that these sounds are part of two different phonemes for Khmer speakers, and these speakers should treat [t] and [t^h] as different sound categories in the context of language processing and comprehension (Bounchan & Moore, 2010).

As this discussion indicates, familiarity with a language changes the way people represent linguistic input. This, in turn, must alter the kinds of learning that people do over that linguistic input. For example, learning that A is associated with X, and B is associated with Y, is likely to be much easier for learners who represent A and B as members of different speech categories than it is for learners who represent A and B as exemplars of the same category. This can be seen, for example, in the context of a statistical learning paradigm. Emberson, Liu and Zevin (2013) demonstrated this using a stream of non-linguistic input with 4 sound categories with 6 exemplars each. These 4 sound categories were grouped together to form "words" made up of two sounds each, presented to participants in a fluent stream. Three of these categories were easy to distinguish from one another, while the fourth was hard to tell apart from the third. When participants performed statistical learning over a stream of two-sound words, they only segmented words that corresponded to the three categories they learned; they could not manage to figure out that another category existed with a better predictive value in the high level organization. This demonstrates the effect that perceptual organization has on our overall perception of speech through artificial stimuli.

In the current experiment, I want to extend this idea of perceptual differences developed through language exposure to a statistical learning paradigm using more realistic linguistic stimuli. Instead of exposure to artificial sound categories, this experiment will use lexical tone as an additional cue for word segmentation on top of the language stream. Tone factors into languages like Thai, Zulu, and Mandarin Chinese; these languages use pitch contours to distinguish between words or to indicate certain grammatical structuring of words. In Mandarin, the most widely spoken tone language, we see a four-tone system (plus an additional neutral mark that indicates a lack of tone). These tones are lexically distinguishing, which can produce some tongue twister sentences like "*mā mà mǎ de má ma"* ("Is mom scolding the horse's hemp?"). We see tone one in the first word, then tone four, three, and two in the following syllables, as well as the neutral interrogative marker */ma/* which ends the sentence.

Infants and adults can apply their statistical learning mechanism to learn tone sequences, which can follow similar tone contours to the linguistic tone sequences used to distinguish words in Mandarin. When presented with a tone language that contained three tones per word, both infants and adults will segment the tone-words out from the language (Saffran et al., 1999). This task, however, is not inherently linguistic in nature, though this is likely to be somewhat influenced by a learners' linguistic background. Klein et al. (2000) found, via a PET scan of Mandarin and English speakers performing a lexical tone discrimination task, that

Mandarin speakers show left hemisphere activation in areas associated with linguistic experience as well as right hemisphere activation, which correlates with regular perception of tones and music. English speakers only showed this right hemisphere activation, which is another indication that exposure to language changes actual perception of linguistic input. These results suggest that prior experiments with English-speaking adults doing tone segmentation (e.g., Saffran et al., 1999; Saffran & Griepentrog, 2001) do not tap into linguistic representations. However, for Mandarin speakers, presenting a statistical learning language with tonal features may tap into linguistic representations.

To test this hypothesis, we will ask both English and Mandarin speakers to perform a statistical learning task over syllabic materials. Unlike traditional statistical learning experiments, however, these syllables will not be presented in a monotone. Instead, they will have a dynamic pitch contour. This pitch contour will be manipulated to create two languages. In one language, each statistically-defined word will have a consistent pitch contour; in the other language, words will occur with a random pitch contour. We predict that Mandarin speakers will do well when the tones are consistently associated with a single word, since this mirrors the pattern they have in their own language. When the tones are randomly assigned to each word, however, we predict that Mandarin speakers will perform poorly. If they store tone as an important factor for each word they hear, they will fail to make a generalization over the words due to the inconsistent tone cues they hear along with them. Monolingual English speakers and speakers of non-tonal languages, however, should not be disturbed by the presence of tone on top of the words in the language. Since pitch contours do not serve as lexically distinguishing factors, these participants should ignore the pitch, regardless of whether it is consistent or inconsistent, and generalize only over the lexical content in the speech stream.

Method

Participants

Fifty-two students aged 18-25 were recruited on the Psychology department's experiment system that requires introductory psychology students to take part in 3 studies each, and also via flyers and announcements provided to the sorority and fraternity communities at Carnegie Mellon University.

These participants received either five dollars or a psychology study credit. Each participant was randomized via coin flip into either the consistent or inconsistent tones condition.

Twenty-two of these participants were Mandarin-English speakers of varying levels of proficiency. Any participant that listed experience with Mandarin Chinese was counted as a tone language speaker. Thirty participants were monolingual English speakers or bilingual to some degree with English and another non-tonal language. All participants were assessed via a verbal fluency task to test their proficiency in their second language. Self-reported monolingual English speakers were tested on the language they learned for a brief period in high school.

Stimuli

This experiment uses two artificially synthesized languages composed of 4 disyllabic (CV.CV) words, as seen in the word table below. The first language, also called the consistent language, matches up one of four specific Chinese pitch contours to each one of the four words. These contours are made up of pairs of the four Chinese tones. Tone 1 is a single high tone. Tone 2 is a rising tone, starting at a medium level and ending high. Tone 3 is a checkmark-shaped contour, starting in the middle, dipping lower and ending up high. Tone 4 is a falling tone, starting high and ending low.

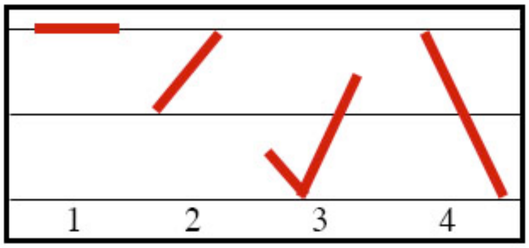

Fig 1. A visual depiction of the tones.

The table below shows the pairs of tones along with an example Chinese word that shares that tone contour.

Table 1. A table of the words used in the language and their consistent tone pairs

Words	Tone Pair Contour	Example Chinese Word
baku	2-1	míngtiān (tomorrow)
tiro	1-3	zhōngwǔ (noon)
pido	4-2	miàntiáo (noodles)
lagu	3-4	hǎokàn (pretty)

The second language, or the inconsistent language, randomly assigns one of the above tone pair contours (2-1, 1-3, 4-2, and 3-4) for every word instance in the language. For example, /lagu/ will appear ¼ of the time with a 2-1 contour, ¼ with 1-3, ¼ with 4-2 and ¼ with 3-4. This creates an inconsistent tonal cue for the listener.

From these words we also created a set of part-words, which are a sequence of syllables that consists of the last syllable of one word and the first syllable of another, like /ku-pi/, taken from baku and pido. These each occur about ¼ as often in the exposure phase than the words. The part-words are used in the test phase to assess how well the participants learned the words of each language. For both the consistent and inconsistent condition, the same part-words were used.

Table 2. A table of the part-words used in the testing phase.

Part-word	Words Used	Pitch Contour
ku-ti	baku + tiro	1-1
ro-pi	tiro + pido	3-4
do-la	pido + lagu	2-3
gu-ba	lagu + baku	4-2

Procedure

Pre-Test

We gathered several pieces of information about each participant in order to determine the extent to which the participant was bilingual and in which languages. The primary test was a verbal fluency task where the participant had 30 seconds to name as many animals as they can, first in their L1 at the beginning of the experiment, and then in their L2 after they finish the statistical learning task. The final result is the ratio between the number of animals in L2 to the number of animals in L1.

We also gathered self-reports of how fluent the participant is at reading and comprehension on a 5 point Likert scale, as well as a yes/no self-report as to whether or not the participant considers himself or herself bilingual.

Segmentation Phase

The exposure phase consisted of a five-minute loop of either the consistent language or the inconsistent language, presented via EPrime. The participant was seated in front of a computer with headphones on, listening quietly to the stimuli. The participant then proceeded to the test trials, which presented them with a forced choice task with sixteen word/part-word pairs. A word is one of the four items used to synthesize the language, and a part-word is the joining of the last syllable of one word with the first syllable of the next, creating a sequence that does occur in the speech stream but only ¼ as much as the words.

We would expect the consistent condition participants who speak a toned language to do well at separating the same-pitch words from the part-words, but to do worse separating the changed-pitch words from the part-words. In contrast, we would expect the consistent condition participants who do not speak a toned language to do equally well separating words from part-words, regardless of the tone given to them. In the inconsistent condition, we would expect toned-language participants to struggle with separation, while the non-toned-language participants to do as well as they did when the tones were consistent.

Results

Each subject's score was recorded as a fraction out of sixteen. A 2x2 ANOVA was performed with Condition (Inconsistent, Consistent) and Toned Language (Yes, No) as between-subjects factors. The interaction between Condition and Tone was significant, ($F(1, 52) = 5.39$, $p<0.05$), so the individual simple main effects were then tested. A Welch two-sample t-test was performed with Language Condition given Tone Language = Yes, and indicated a significant result ($t(22) = 2.81$, $p<0.05$). This shows that Mandarin speakers are performing better in the consistent language condition ($M = .83$) than the inconsistent language condition (M = .64). In contrast, the t-test for Language Condition given Tone Language = No was not significant, so the non-tonal language speakers are performing the same in both language conditions (consistent $M = .71$; inconsistent $M = .76$). There was also a significant result in Tone Language given Language Condition = Consistent ($t(27) = 2.18$, p<.0.05). This indicates that the Chinese speakers ($M = .84$) are performing significantly better than the non-tonal language speakers ($M = .71$), when both groups are presented with linguistic stimuli in which tones are consistently associated with particular words.

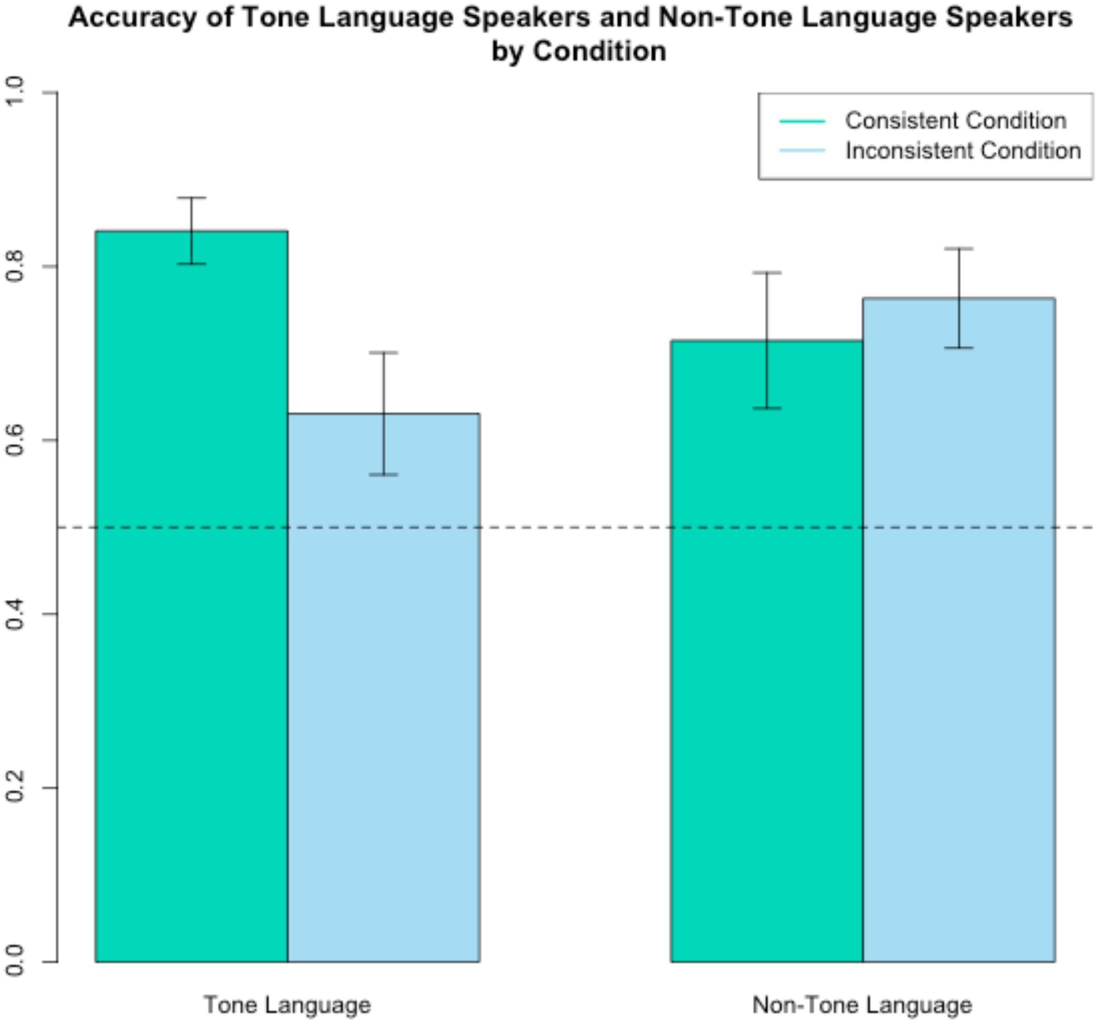

Fig 2. A bar chart of the means in each condition

Discussion

These results indicate that speakers of Mandarin perform significantly better at segmenting a speech stream when provided with a consistent tonal cue to each word rather than a random inconsistent tonal cue, while non-tonal language speakers, who perform the same in both conditions, do not utilize the tonal cue for segmentation.

One potential explanation for these results would be that the addition of a consistent tone cue to the stimuli language provided an additional segmentation cue to the Mandarin speakers. This would cause the Mandarin speakers to perform better in the consistent language condition than the inconsistent condition, which is what we observe, and leave the non-tonal language speakers performing similarly in both conditions, since they do not use tone as a cue for segmentation. The transitional probabilities between the syllables in the language, however, provide the exact same information as the transitions between the tones. It is possible that the five-minute exposure phase was too short to get completely accurate information via syllable transitions, and the tones facilitated in the process. This has been seen with computer models that use both transitional probabilities between syllables and lexical stress to segment child-directed speech (Christiansen et al., 1998). In that case, it would make sense that the combination of the two cues led to increased performance.

It is also interesting to note that the Mandarin speakers still performed above chance in the inconsistent tonal condition, where the tonal cue was designed such that the Mandarin speakers would hear sixteen distinct words: every combination of the four words in the language with the four Mandarin tones. It would be valuable to see if the Mandarin speakers would be hindered further if the tones were not selected from the four Mandarin tones, or if the tones violated certain grammatical rules of tone, known as *tone sandhi.* For example, the third tone cannot be followed by another third tone in the same word. If this occurs, the first third tone will change to a second tone. Across a word boundary, however, this rule does not apply (Zhang, 1988). As such, Mandarin speakers may segment out the wrong words if the cue was actually counter to the grammatical rules in their language.

To further confirm that the non-tonal language speakers are not representing tone, a future experiment is planned with the same exposure and testing paradigm as this experiment. The test items, however, are switched out with monotone versions of the words. As a monolingual English speakers or a non-tonal language speaker, the tonal contour over the words should not affect the overall processing of the word, since the word's meaning doesn't depend on its pitch. For Mandarin speakers, however, this should drastically reduce their consistent tonal language performance, since the test items are effectively new words that the speakers have not experienced before, given that tone is lexically distinguishing for them.

Further exploration could expand upon the research done by Best et al. (1987) and explore statistical learning with words that contain clicks from the Bantu language Zulu. It is not entirely accurate to say that tones do not help distinguish information in English, since a downward tonal contour at the end of a sentence indicates a statement, while an upward contour implies a question. These correspond with Tone 4 and Tone 2 in Mandarin, respectively (Shen, 1984). This use of tone as a grammatical distinguisher may effect the level to which speakers of English or romance languages treat tone when learning words. A language involving an item that is lexically important for Zulu speakers but does not even provide exist in allophonic free variation in English would help to understand whether the mere fact that English speakers can distinguish between clicks means that they would represent them as different items in words when parsing through a speech stream.

These experimental results indicate that statistical learning is not just performed over the external auditory information, but more importantly is done over our internal representations of this input. The English and non-tonal language speakers perceive the same stimuli with tone contours as the Mandarin speakers, yet represent the words without the tone contours. In the consistent condition, these contours provide an additional helpful cue for segmentation, yet the English speakers do not use it. This can be explained by the English speakers' internal representations of the words, shaped by their exposure to a non-tonal language, that ignore these tone contours. This account is consistent with infants' phonetic development, as they focus on their own language's phonemes around 12 months and begin to group together sounds that are not lexically distinguishing, like aspirated and unaspirated p for English speakers, (e.g., Werker & Tees, 1984) or in this case, various tonal contours that could naturally occur over words. Adults then continue to have difficulty with discrimination between items that they grouped together as allophones or, in this case, allotones.

References

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1987). Examination of perceptual reorganization for non-native speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Status Report on Speech Research,* 2-29.

Bounchan, S., & Moore, S. H. (2010). Khmer Learner English: A Teacher's Guide to Khmer L1 Interference. *Language Education in Asia*, 1(1), 112-123.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes,* 13(2/3), 221-268.

Curtin, S., Mintz, T. H., Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, Vol. 96, 233-262.

Emberson, L. L., Liu, R., & Zevin, J. D. (2013). Is statistical learning constrained by lower level perceptual organization? *Cognition,* 128(1), 82-102.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255, 606-608.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology*, 26(5), 1570-1582.

Jusczyk, P. W., & Aslin, R. N. (1995). Infant detection of the sound patterns of words in fluent speech. *Cognitive Psychology,* 29, 1-23.

Klein, D., Zatorre, R. J., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *Neuroimage,* 13(4), 646-653.

Lew-Williams, C., Saffran, J. R. (2012). All words are not created equal: expectations about word length guide infant statistical learning. *Cognition*, 122(2), 241-6.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition,* 82, B101-B111.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning in 8-month-old infants. *Science,* 274(5294), 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27-52.

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology,* 37(1), 74-85.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, 12(4), 112-114.

Shen, X.S. (1989). Toward a register approach in teaching Mandarin tones. Journal of Chinese Language Teachers Association, 24, 27-47.

Singh, L. (2008) Influences of high and low variability on infant word recognition. *Cognition*, 106, 833-870.

Singh, L., White, K, S. & Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, 4, 157 - 178.

Thiessen, E.D., & Saffran, J.R. (2003). When cues collide: Use of statistical and stress cues to word boundaries by 7- and 9-month-old infants. *Developmental Psychology*, 39, 706-716.

Traeger, G. L. (1942). The phoneme "T": A study in theory and method. *American Speech*, 17(3), 144-148.

Werker, J. F., & Tees, R. C. (1894). Cross-language speech perception: Evidence for perceptual reorganization in the first year of life. *Infant Behavior and Development,* 7, 49-63.

Zhang, Zheng-sheng. 1988 Ph.D. diss. Tone and tone sandhi in Chinese. (Advisor: Arnold Zwicky, Linguistics; DEALL Committee Member: Marjorie K.M. Chan)