Can voice conversion be used to reduce non-native accents?

Sandesh Aryal, Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University, USA

{sandesh,rgutier}@cse.tamu.edu

Abstract

Voice-conversion (VC) techniques aim to transform utterances from a source speaker to sound as if they had been produced by a target speaker. This includes not only organic properties (i.e., voice quality) but also linguistic cues (i.e., regional accents) of the target speaker. For this reason, VC is generally ill-suited for accent-conversion (AC) purposes, where the goal is to capture the voice quality of the target speaker but the regional accent of the source speaker. In this paper, we propose a modification of the conventional training process for VC that allows it to perform as an AC transform. The approach consists of pairing source and target vectors based not on their ordering within a parallel corpus, as is commonly done in VC, but based on their linguistic similarity. We validate the AC approach on a corpus containing native-accented and Spanish-accented utterances, and compare it against conventional VC through a series of perceptual listening tests. We also analyze the extent to which phonological differences between the two languages (Spanish and American English) help predict the relative performance of the two methods.

Keywords: Voice conversion, accent conversion, non-native speech, Spanish accent.

1. Introduction

During the last two decades, a few studies have suggested that it would be beneficial for second language (L2) learners to be able to listen to their own voices producing native-accented speech [1-3] The rationale behind this proposal is that removing information that is only related to the teacher's voice quality makes it easier for students to perceive differences between their accented utterances and their ideal accent-free counterparts. As a step towards this goal, in the past we have explored a number of "accent-conversion" techniques to transform the voice of an L2 learner in a way that resembles native pronunciation, both using vocoding [1] and articulatory unit-selection synthesis [2]. Though both approaches can reduce accents they are not without problems; vocoding leads to the perception of a third speaker (one that is neither the L1 nor the L2 speaker), and unit-selection synthesis requires access to a very large acoustic-articulatory corpus, which is impractical in most learning scenarios.

A closely related problem to accent conversion is that of voice conversion. In voice conversion, one seeks to transform utterances from a source speaker so they sound as if they had been produced by a different (but known) target speaker. To train a voice-conversion mapping, one has to generate a training set containing pairs $(\boldsymbol{x_i}, \boldsymbol{y_i})$ of spectral feature vectors, a vector $\boldsymbol{x_i}$ (generally MFCCs) for the source speaker, and the corresponding vector $\boldsymbol{y_i}$ for the target speaker. The pairing is generally accomplished by force-aligning parallel recordings from both speakers. This step is not only one of the major limitations of the conventional voice-conversion process, but also the main reason why it cannot work for the more specialized problem of accent conversion: if the source speaker is native-accented and the target speaker is foreign-accented, voice conversion will preserve not only the voice quality of the target but also his or her accent.

In this paper we show how a modification to the above training process for voice conversion can yield appreciable reductions in the perceived foreign accent of the target voice. The approach consists of pairing source and target vectors based not on their ordering within the corpus, but on their linguistic similarity. Namely, we pair each frame $\boldsymbol{x_i}$ in the source speaker corpus with the closest frame $\boldsymbol{y_j^*}$ in the target speaker corpus following vocal tract length normalization (VTLN). Likewise, we pair each VTLN frame $\boldsymbol{y_i}$ in the target speaker corpus with the closest frame $\boldsymbol{x_j^*}$ in the source speaker corpus. The VTLN step is critical to ensure that the pairing is based on linguistic content rather than on acoustic similarity.

The remainder of this paper is structured as follows. Section 2 reviews previous work in speech processing methods for accent conversion. Section 3 describes the proposed accent conversion method and the baseline voice conversion technique upon which it is based. Section 4 describes the experimental protocol used to evaluate the effectiveness of the method on a Spanish speaker of American English. Results from perceptual listening tests are included in section 5, along with an analysis of correlation with differences between the Spanish and English phonetic inventory. Finally, section 6 concludes with a discussion of results and directions for future work.

2. Related work

Various approaches have been explored to apply linguistic gestures from a native speaker to a voice quality carrier from a foreign speaker [1, 2, 4-6]. In an early study [5], Yan et al. used HMMs to track formant of different vowels in three major English accents (British, Australian and general American), and then performed spectral transformation between the source and target accents using formant ratios learned from the data. The authors also modified pitch trajectory and duration with TD-PSOLA [7] using rules learned from pitch and duration analysis of vowels in those accent groups. A group of British listeners evaluated the accent conversion method and found that the Australian-to-British accent conversion was 78% accurate and British-to-American accent conversion was 71% accurate.

Felps et al. [1] extended the segmental modification method so that it could be applied not only to vowels and diphthongs but also to unvoiced phones. The authors replaced the spectral envelope of utterances from an L2 speaker with that of an L1 speaker (following vocal tract length normalization) and modified prosody with FD-PSOLA [7]. Subjective evaluation showed a significant reduction in accent (from an initial rating of 5 down to 2) but the resulting synthesis sounded like a third speaker.

In more recent work, Felps et al. [2] used articulatory information to represent linguistic gestures for the purposes of accent conversion. The authors developed a unit-selection synthesis where accent conversion was achieved by selecting L2 units with similar articulatory gestures as those from a target L1 utterance. The method was successful in maintaining the identity of the foreign speaker but the accuracy of the accent reductions and the overall synthesis quality were limited by the small size of the articulatory corpus.

3. Methods

In contrast with these previous studies, here we propose a GMM-based voice conversion technique as the basis for accent conversion [8]. Unlike conventional voice conversion, though, where source and acoustic vectors are matched using forced alignment in a parallel corpus, our approach consists of pairing source and target vectors based on their acoustic similarity following VTLN. Both approaches are illustrated in Fig. 1.

3.1 Pairing acoustic vectors

The first step in our AC method is to apply VTLN in order to minimize physiological differences in the vocal tract of the two speakers. For this purpose, we use dynamic time warping to align parallel utterances from the L1 and L2 speakers, each utterance represented as a sequence of 24 MFCCs. Following Panchapagesan and Alwan [9], we then learn a linear transform between the MFCCs of both speakers using least squares:

$$W = \arg\min \|\boldsymbol{x} - W\boldsymbol{y}\|^2 \qquad (1)$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of MFCC parameters from the L1 and L2 speakers, respectively, and W is the VTLN transform. Next, for each L1 vector $\boldsymbol{x_i}$ we find its closest L2 vector $\boldsymbol{y_j^*}$ as:

$$\boldsymbol{y_j^*} = \arg\min_{\forall \boldsymbol{y}} \|\boldsymbol{x_i} - W\boldsymbol{y}\|^2 \qquad (2)$$

To make the search for the closest frame more efficient, we first group all L2 acoustic frames into 512 clusters using *k*-means. Then, for each L1 frame $\boldsymbol{x_i}$, we first find the closest L2 cluster and then the closest frame from those within that cluster. Likewise, we repeat the process for each L2 vector $\boldsymbol{y_i}$ to find its closest match $\boldsymbol{x_j^*}$:

$$\boldsymbol{x_j^*} = \arg\min_{\forall \boldsymbol{x}} \|\boldsymbol{x} - W\boldsymbol{y_i}\|^2 \qquad (3)$$

This results in a lookup table where each L1 and L2 vector in the database is paired with the closest vector from the other speaker. It is this lookup table that we then use to train a GMM, as explained next.

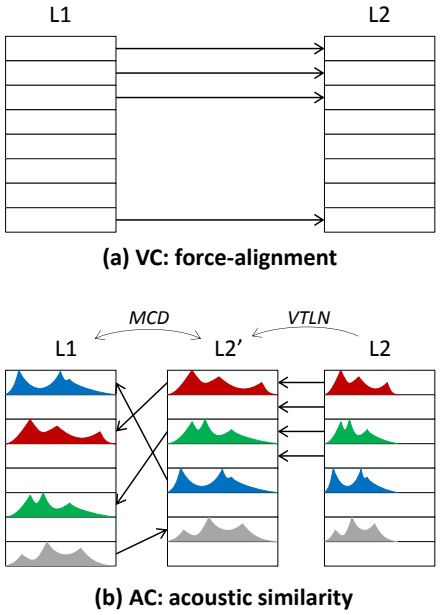

Fig. 1. (a) Conventional approach to voice conversion. (b) Our approach to accent conversion. MCD: Mel Cepstral Distortion.

3.2 Baseline voice conversion method

Following Toda et al. [10], we use a GMM with global variance [8] to learn an acoustic mapping between the two speakers. Let $\boldsymbol{x_t}$ be a vector of static and dynamic (delta) MFCCs for the L1 speaker at frame t, and $\boldsymbol{y_t}$ be the corresponding vector for the L2 speaker. Then, we model the joint distribution $\boldsymbol{z}_t = [\boldsymbol{x}_t, \boldsymbol{y}_t]$ as

$$P(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \qquad (4)$$

where $\boldsymbol{\lambda}^{(z)} = \{\alpha_m, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\}$ are the GMM parameters (weight, mean and covariance of the m^{th} mixture component, respectively), learned from a training set of joint vectors $\boldsymbol{z}_t$ using expectation-maximization (EM).

Given a trained GMM, we calculate the maximum likelihood estimate of acoustic features considering the dynamics and the global variance (GV) as follows. Let $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2} \ldots]$ denote the sequence of L1 acoustic vectors in a source sentence and $\boldsymbol{Y} = [\boldsymbol{Y_1}, \boldsymbol{Y_2} \ldots]$ the GMM sequence of L2 acoustic vectors, where $\boldsymbol{X}_t = [\boldsymbol{x}_t, \boldsymbol{\Delta x}_t]$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t, \Delta \boldsymbol{y}_t]$. Consider also the within-sentence variance of the d^{th} acoustic feature $y_t(d)$ given by $v(d) = E[(y_t(d) - E[y_t(d)])^2]$. Thus, the GV of the static acoustic feature is written as $\boldsymbol{v}(\boldsymbol{y}) = [\, v(1), v(2) \ldots v(D)]$ where D is the dimension of $\boldsymbol{y}_t$, and $\boldsymbol{y}$ is the sequence $[\boldsymbol{y}_1, \boldsymbol{y}_2 \ldots]$. Now, the time sequence of estimated acoustic vectors (static only) is given by:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)}) \qquad (5)$$

where $\boldsymbol{\lambda}^{(v)} = \{\boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\}$, $\boldsymbol{\mu}^{(v)}$ is the vector of average variance for all acoustic features and $\boldsymbol{\Sigma}^{(vv)}$ is the corresponding

covariance matrix, learned from the distribution of $\boldsymbol{v}(\boldsymbol{y})$ in the training set. The likelihood $P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(z)},\boldsymbol{\lambda}^{(v)})$ is computed as

$$P\left(\boldsymbol{Y}\middle|\boldsymbol{X},\boldsymbol{\lambda}^{(z)},\boldsymbol{\lambda}^{(v)}\right) = P\left(\boldsymbol{Y}\middle|\boldsymbol{X},\boldsymbol{\lambda}^{(z)},\boldsymbol{\lambda}^{(v)}\right)^{w}.\ P\left(\boldsymbol{v}(\boldsymbol{y})\middle|\boldsymbol{\lambda}^{(v)}\right) \quad (6)$$

The distribution of GV, $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$, is modeled by a single Gaussian $\mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y});\boldsymbol{\mu}^{(v)},\boldsymbol{\Sigma}^{(vv)}\right)$. The power term $w\ (=1/2L)$ in equation (6) controls the balance between the two likelihoods. Following [8], we use EM to solve for $\hat{\boldsymbol{y}}$ in equation (5).

3.3 Prosody modification

As shown in previous studies [1, 4, 5], prosody modification is an important part of accent conversion. Following [8], we use the aperiodicity and pitch trajectory of the source (L1) speaker, which captures the native intonation pattern, but normalize it to the pitch range of the target (L2) speaker to preserve his or her natural vocal range. More specifically, given an L1 pitch trajectory $f_1(t)$, we generate the modified L2 pitch trajectory $f_2(t)$ as:

$$\log\left(f_2(t)\right) = [\log(f_1(t)) - \mu_1]\frac{\sigma_2}{\sigma_1} + \mu_2 \quad (7)$$

where (μ_1,σ_1) and (μ_2,σ_2) are the mean and standard deviation of log-scaled pitch of the L1 and L2 speakers, respectively, calculated from the training corpus.

4. Experimental

4.1 Conversion from non-native to native accent (L1→L2)

To test the effectiveness of our AC approach, we compared it against a baseline VC model and utterances from the L1 and L2 speakers, which served as controls. The baseline VC model was similar to the AC model except that the GMM model was trained on DTW-aligned pairs from source and target speakers –see Fig. 1(a), whereas the AC model was trained on acoustically-matched pairs as described in section 3.1. In both cases, the GMM consisted of 128 Gaussian components.

We performed two sets of perceptual listening tests:

1) *Perceived accent*: subjects listened to pairs of utterances (AC-VC, AC-L2, VC-L2) and were asked to select the utterance that sounded the least accented. Order of presentation in the pairs was randomized within subjects.

2) *Perceived speaker identity*: subjects listened to three utterances (A,B,X) and were asked to select whether the speaker in utterance X sounded closer to the identity in A or B. Utterances in X were either AC or VC; utterances A, B were either L1 or L2 (order of presentation was randomized within subjects). Following [1], utterances were played backward.

To ensure that the loss of quality in the AC and VC methods due to the MFCC compression step did not affect the perceptual ratings, control utterances from the L1 and L2 speaker were compressed to MFCC and then resynthesized as described in [11].

Table 1. Summary of the six synthesis models (AP: aperiodicity from STRAIGHT; DTW: dynamic time warping)

Synthesis model	Frame pairing	Source MFCC	AP	Target MFCC
AC12	Acoustic	L1	L1	L2
VC12	DTW	L1	L1	L2
AC21	Acoustic	L2	L2	L1
VC21	DTW	L2	L2	L1
L1	-	L1	L1	L1
L2	-	L2	L2	L2

4.2 Conversion from native to non-native accent (L2→L1)

We also tested the effectiveness of our AC approach to map accents in the opposite direction, i.e., imparting a non-native accent to the voice of a native speaker. For this purpose, we trained AC and VC models in a manner similar as in section 4.1, except we used the L2 speaker as the source speaker, and the L1 speaker as the target speaker. The six types of synthesis evaluated are summarized in Table 1.

4.3 Corpus

The speech corpus consisted of parallel recordings from a non-native speaker (whose first language was Spanish) and a native speaker of American English, previously described in [2]. Both subjects recorded the same 344 sentences chosen from the Glasgow Herald corpus. In addition, L2 recorded 305 sentences not spoken by L1. Out of the 344 sentences shared among both speakers, we randomly selected 294 sentences to train the GMM, and saved the remaining 50 sentences for testing purposes. For each sentence, we computed 25 MFCCs ($MFCC_{0:}$ energy; $MFCC_{1\text{-}24:}$ spectral envelope) as well as *pitch* and *aperiodicity* from the STRAIGHT [12] spectrum sampled at interval of 5ms[1].

5. Results

Listening tests were performed on Amazon's Mechanical Turk. Following [2], to qualify participants first had to pass a screening test that consisted of identifying various American English accents. Participants were also asked to list their native language/dialect and any other fluent languages that they spoke. If a subject was not a monolingual speaker of American English then their responses were excluded from the results. Participants were paid $2 for completing the test.

5.1 L1→L2 conversion

Thirteen participants rated the accent and identity of the AC12 and VC12 models on a set of 12 sentences, randomly selected from the 50 sentences in the test set. Both models were perceived to be less foreign-accented than the original L2 utterances. On average, listeners found VC12 to be less accented than the original L2 utterances 91% of the times (std 9%). Likewise, listeners found AC12 less foreign-accented than L2 86% of the times (std 8%). This result would suggest

[1] STRAIGHT was also used to resynthesize utterances from the output of the GMM-GV model.

that conventional voice conversion (VC12) was more effective in reducing accents than our proposed method (AC12). However, when both models were compared against each other, participants found AC12 to be less accented than VC12 59% of the times (std 12%). Finally, results from the ABX identity test show that participants found AC12 closer to L2 than to L1 78% of the times on average.

In summary, these results indicate that the proposed AC method is more effective in reducing accent than conventional VC, while at the same time it preserves the identity of the L2 speaker.

5.2 L2→L1 conversion

Twelve participants rated the accent and identity of the AC21 and VC21 models on a set of 12 sentences, randomly selected from the 50 sentences in the test set. Both models were perceived to be more foreign-accented than the original L1 utterances. On average, VC21 was rated as more foreign-accented than L1 (mean 97%; std 9%), and AC21 was rated as more foreign-accented than L1 as well (mean 96%; std 9%). More importantly, when compared against each other AC21 was rated as more foreign-accented than VC21 (mean 66%, std 16%). Finally, the voice quality of AC21 was found more similar to L1 than to L2 (mean 65%; std 28%).

In summary, these results show that the proposed AC method is also more effective than the baseline VC method in imparting a non-native accent to a native speaker, while it also preserves the identity of the L1 speaker.

5.3 Correlation with differences in the L1 and L2 phonetic inventories

As a final step, we analyzed whether the effectiveness of the AC model could be explained from differences in the phonetic inventory of the two languages [13-15]. In particular, the English language includes a number of consonants that do not exist in Spanish, most significantly the fricatives [v], [z], [θ], [ʃ], [ʒ] and [ð], the affricate [ʤ], the pseudo-fricative [h], and the liquid [ɹ]. Spanish also does not have lax vowels, the schwa as well as r-colored vowels.

Thus, for each sentence in the listening tests we computed the number of phonemes that did not exist in Spanish ($N_{p \notin L2}$), our rationale being that the larger this number the more difficult it would be for the L2 speaker to pronounce the sentence. Then, we computed the correlation coefficient between $N_{p \notin L2}$ and the proportion of listeners who found the AC12 synthesis less accented than the VC12 synthesis. Results reveal a very strong correlation (0.86) between both measures, which indicates that the benefits of the AC method are more significant for sentences that are harder to produce by the L2 speaker. We also computed the correlation between $N_{p \notin L2}$ and the proportion of listeners who found AC12 less accented than L2; in this case, the correlation was 0.63, which adds further support to the previous conclusion. In contrast, the performance of baseline voice conversion method appears to be unrelated (correlation 0.08) to the difficulty of the test sentence.

6. Discussion

We have presented a speech modification method that can be used to transform L2 utterances to sound more native-accented. The method is based on conventional GMM techniques for voice conversion, but uses a different strategy to match frames from the source (L1) and target (L2) speakers. Namely, we apply vocal tract length normalization and then perform a bidirectional match between frames of the two speakers using Mel Cepstral Distortion as a measure of similarity; the resulting lookup table of source-target vectors is then used to train a GMM.

To test the effectiveness of our method, we compared it against a baseline voice-conversion model trained on DTW-aligned pairs of source-target utterances. Listening tests show that our accent conversion method can transfer the accent of the source speaker more effectively than voice conversion, regardless of the direction in which the transformations are applied, i.e., making L2 utterances less foreign-accented vs. making L1 utterances more foreign-accented.

Our results also show that the accent conversion method is most beneficial when used on utterances that are difficult to produce by L2 speakers, as measured by the number of phones in the utterance that do not exist in the L2 phonetic inventory. Further insights may be obtained by analyzing phonotactic differences between the two languages. A classic example in Spanish is the lack of word-initial clusters that begin with /s/; in these cases, Spanish speakers tend to produce such words (e.g., star, scar, small, Spain) with an initial /e/. One may also consider whether the particular error has high or low functional load; as an example, contrast between initial p/b has a high relative functional load, whereas final t/d has a lower functional load [16].

The L2 speaker in our study had lived in the United States for 16 years at the time of the recordings so, while he had a non-native accent, he is functionally bilingual. Further work is required to assess the effectiveness of the proposed method as a function of the proficiency or experience of the L2 speaker. Likewise, additional work is required to test the effectiveness of our method on L2 speakers with different L1 (e.g., Spanish, Chinese, German).

Further improvements in the accent conversion model may also be obtained by imposing constraints on the pairing of acoustic vectors in section 3.1. As an example, one may eliminate source-target pairs that have high Mel Cepstral Distortion. Likewise, performance may also be improved by considering additional information when matching source-target pairs, such as dynamic features (delta and delta-delta), features from the STRAIGHT aperiodicity spectrum, or linguistic features predicted from speech acoustics such as sound classes (e.g., place and manner of articulation).

7. Acknowledgments

This work is supported by NSF award 0713205. We are grateful to Prof. Steve Renals and the Scottish Informatics and Computer Science Alliance (SICSA) for their support during RGO's sabbatical stay at CSTR (University of Edinburgh).

References

[1] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech communication*, vol. 51, pp. 920-932, 2009.

[2] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2301-2312, 2012.

[3] D. Felps and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1030-1040, 2010.

[4] M. Huckvale and K. Yanagisawa, "Spoken language conversion with accent morphing," in *Proc. ISCA Speech Synthesis Workshop, Bonn, Germany*, 2007, pp. 64-70.

[5] Q. Yan, S. Vaseghi, D. Rentzos, and C.-H. Ho, "Analysis and synthesis of formant spaces of British, Australian, and American accents," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 676-689, 2007.

[6] N. Campbell, "Foreign-language speech synthesis," *Proceedings SSW3*, pp. 177-180, 1998.

[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453-467, 1990.

[8] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2222-2235, 2007.

[9] S. Panchapagesan and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," *Computer speech & language*, vol. 23, pp. 42-64, 2009.

[10] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, pp. 215-227, 2008.

[11] S. Aryal and R. Gutierrez-Osuna, "Articulatory inversion and synthesis: towards articulatory-based modification of speech," *Accepted to appear in: ICASSP*, 2013.

[12] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *ICASSP*, 1997, pp. 1303-1306.

[13] B. Goldstein, "Transcription of Spanish and Spanish-influenced English," *Communication Disorders Quarterly*, vol. 23, pp. 54-60, 2001.

[14] L. A. Helman, "Building on the sound system of Spanish: Insights from the alphabetic spellings of English-language learners," *The Reading Teacher*, pp. 452-460, 2004.

[15] H. You, A. Alwan, A. Kazemzadeh, and S. Narayanan, "Pronunciation variations of Spanish-accented English spoken by young children," in *Interspeech*, 2005, pp. 749-752.

[16] G. Jesse, "Beaches and peaches: Common pronunciation errors among L1 Spanish speakers of English," in *PSLLT*, 2012, pp. 205-215.