Generating segmental foreign accent

María Luisa García Lecumberri[1], *Roberto Barra-Chicote*[2], *Rubén Pérez Ramón*[1], *Junichi Yamagishi*[3,4], *Martin Cooke*[5,1]

[1]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[2]Grupo de Tecnología del Habla, Universidad Politécnica de Madrid, Madrid, Spain
[3]Center for Speech Technology Research, University of Edinburgh, UK
[4]National Institute of Informatics, Japan
[5]Ikerbasque (Basque Foundation for Science), Bilbao, Spain

garcia.lecumberri@ehu.es

Abstract

For most of us, speaking in a non-native language involves deviating to some extent from native pronunciation norms. However, the detailed basis for foreign accent (FA) remains elusive, in part due to methodological challenges in isolating segmental from suprasegmental factors. The current study examines the role of segmental features in conveying FA through the use of a generative approach in which accent is localised to single consonantal segments. Three techniques are evaluated: the first requires a highly-proficiency bilingual to produce words with isolated accented segments; the second uses cross-splicing of context-dependent consonants from the non-native language into native words; the third employs hidden Markov model synthesis to blend voice models for both languages. Using English and Spanish as the native/non-native languages respectively, listener cohorts from both languages identified words and rated their degree of FA. All techniques were capable of generating accented words, but to differing degrees. Naturally-produced speech led to the strongest FA ratings and synthetic speech the weakest, which we interpret as the outcome of over-smoothing. Nevertheless, the flexibility offered by synthesising localised accent encourages further development of the method.

Index Terms: Foreign accent, speech synthesis, splicing

1. Introduction

Foreign accent has important effects on communication but little is known about the effect of specific FA characteristics since many studies use holistic evaluations of speakers' pronunciations. What research there has been on individual cues has mostly focused on how suprasegmental characteristics such as nuclear stress [1], syllable patterns [2], duration [3], speech rate [4] and intonation and pauses [5] affect intelligibility and degree of foreign accent (DFA). Low-pass filtering techniques and content-masked speech [6] have also been used to explore the role of non-segmental factors in FA.

Another method involves correlating accent judgements with error types. One study correlated DFA, intelligibility, and comprehensibility with 'accent features' – grammatical errors, phonemic errors, prosody and speaking rate – as coded by experimenters [7]. Listeners also indicated which aspects of accent were most noticeable. Interestingly, these open responses indicated segmental errors as the phonetic cues most strongly reflecting FA. The impact of consonants' functional load on FA has also been investigated [8], with the finding that functional load errors were reported in the DFA but had a smaller impact on comprehensibility. Correlational methods represent an initial approach to isolating those speech aspects that may be interacting with FA measures, but they do not provide very detailed and direct information about the effect of specific speech characteristics on FA ratings since individual phonetic cues are not controlled.

The current study adopts a different approach to the isolation of segmental-level carriers of FA in consonants. Using speech material from an advanced bilingual talker, three different *generative* methods for foreign accent are evaluated:

1. The NATURAL condition uses bilingual 'code-switching' at the segmental level i.e., introducing a single accented phoneme into an otherwise unaccented stimulus during natural speech production (e.g., /xaʊs/ instead of /haʊs/).
2. The SPLICED condition involves replacing certain English consonants with Spanish consonants (e.g., Spanish /x/ replacing English /h/ in /haʊs/).
3. The SYNTHETIC approach uses model-based combination of two synthetic voices, one for each of English and Spanish, trained using speech from the bilingual talker (e.g., learnt context-dependent features for Spanish /x/ used in sequence with similar features for English /aʊs/ prior to speech generation).

The three approaches differ in their benefits and potential drawbacks. The NATURAL condition maintains speaker consistency and is free of processing artefacts, but segment-level code-switching relies on the metalinguistic ability of a speaker to control speech articulation, which may be challenging for some accented segments. Splicing does not require linguistically-sophisticated speakers but may lack credible voice continuity and be prone to artefacts. Arguably, text-to-speech synthesis (TTS) provides the most flexible alternative since, once voice models are learnt, arbitrary sequences can be generated on demand. Some of the best TTS systems are based on unit selection – essentially a sophisticated form of splicing – but the advent of HMM-based TTS methods (HTS) provides additional potential such as the possibility to control not just the presence of accented segments but the degree of their accentedness.

The goal of the current study is to assess the merits of the three techniques in generating segmental foreign accent. Cohorts of native (N; English) and non-native (NN; Spanish) listeners both identified and rated the degree of foreign accent of

Copyright © 2014 ISCA

Spanish-accented English words produced using the three methods, alongside non-accented control tokens.

2. Generating accented words

This section explains the process of generating accented words for the three processing techniques, NATURAL, SPLICED and SYNTHETIC. For each approach a non-accented version was also constructed, resulting in 6 sets of stimuli.

2.1. Bilingual speech material

All techniques used speech data from a highly-competent female bilingual speaker of English and Spanish who showed no trace of foreign accent in either of the two languages as judged by native listeners. It is worth noting that speakers meeting this very strict definition of balanced bilingual are extremely rare. Several corpora were collected from this talker:

- A large corpus of read sentences in Spanish (around 2 hours of speech) from the phonetically-balanced Albaycin corpus [9] as well as newspaper sentences.
- An equivalent English corpus of read sentences extracted from newspapers [10].
- A custom-designed corpus of 108 English words containing consonants known to be problematic for Spanish speakers of English. These consonants, in initial and in some cases medial position, are listed in Table 1.
- The same corpus of English words with designated target Spanish segments interleaved by the speaker (i.e., phonetic code-switching).
- A custom-designed corpus of Spanish nonsense words (e.g., /xasa/) containing consonants in pre- and intervocalic contexts required for splicing accented segments (e.g., /x/ and part of the following /a/ replacing the initial part of /haʊs/).

Table 1: *Problematic English consonants and some realisations typical of Spanish learners.*

	Realisation	Examples
/h/	velar/uvular	/haʊs/ ↦ /xaʊs/
/j/	affricate /dʒ/	/jes/ ↦ /dʒes/
/j/	fricative /ʝ/	/jes/ ↦ /ʝes/
/k,t/	lack of aspiration	/k^{h}əʊld/ ↦ /kəʊld/
/ɹ/	trill	/ɹʌn/ ↦ /rʌn/
/ɹ/	tap	/veɹi/ ↦ /veri/
/v/	/b/	/veɪn/ ↦ /beɪn/
/v/	approximant /β̞/	/ɹɪvɜ:b/ ↦ /ɹɪβ̞ɜ:b/
/ð/	/d/	/ðɪs/ ↦ /dɪs/
/ð/	approximant /ð̞/	/mʌðə/ ↦ /mʌð̞ə/
/w/	/ɡw/	/weɪn/ ↦ /ɡweɪn/
/dʒ/	/j/	/dʒu:s/ ↦ /ju:s/

Speech material was collected in a recording studio in the Phonetics Laboratory at University of the Basque Country using an AKG 4500 table mike and RME A-D converter. Recording sessions were split according to language of the recording to avoid phonetic code-switching except in the condition where that was required. For the same reason, the speaker was given oral and written instructions in the single language of the recording session and produced several repetitions of the word materials from which the final corpus of 108 word exemplars was selected.

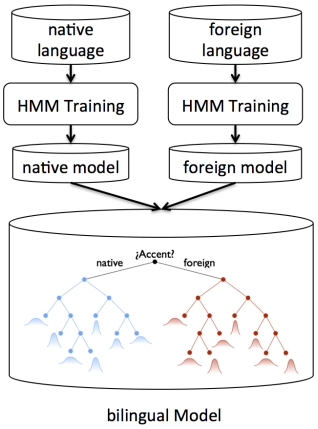

Figure 1: *Voice building procedure for the TTS voices*

2.2. Natural speech

The non-accented NATURAL speech condition simply consisted of isolated English words. For the accented versions, the talker was presented with aural examples of the required consonantal substitution and attempted to produce the English word with the Spanish target consonant. This process was repeated until adequate exemplars were produced.

2.3. Splicing

In the SPLICED condition, target Spanish consonants along with a portion of the surrounding sound(s) were extracted from Spanish nonsense tokens, replacing the non-accented consonant. Splicing was a semi-automatic process which made use of manual phonemic segmentations of the English words and Spanish nonsense tokens. A Praat [11] script used overlap-add with 50 ms overlap to replace the target consonant. For cases where the preceding or following sound was too short (mainly schwas or occlusive consonants) a reduced overlap duration was used. Following auditory screening, two further manual adjustments were carried out. First, where the intensity of the inserted sound differed by more than 10 dB from the original, its level was adjusted to reduce the difference, using the "Modify scale intensity" method of Praat. Second, in some cases an abrupt difference in F0 led to perceptual streaming of the inserted sound. In these cases the F0 contour was smoothed in Praat. To quantify the effect of any artefacts introduced by splicing, a non-accented spliced condition was created by splicing of the target consonant from a different English exemplar of the same word (from the same talker) using the process described above.

2.4. Synthesis

A bilingual synthetic voice was built based on the combination of two language-independent synthetic voices trained using English and Spanish data from the bilingual talker, as outlined in Figure 1. The individual synthetic voices were built using statistical parametric speech synthesis via the HTS Toolkit [12] adapted for English [13] and Spanish [14]. The HMM-based speech synthesis system involves three processes: speech analysis, HMM training, and speech generation.

In the speech analysis part, three kinds of parameters for the STRAIGHT [15] mel-cepstral vocoder with mixed excitation (mel-cepstrum, log $F0$ and a set of aperiodicity measures) were extracted as HMM feature vectors. Context-dependent multi-

stream left-to-right Multi-Space Distribution Hidden Semi-Markov Models (MSD-HSMMs, [16]) were trained using a maximum likelihood criterion. For speech output, acoustic feature parameters were generated from the MSD-HSMMs using a parameter generation algorithm [17]. In order to avoid the muffled effect of HMM-synthesised speech, the system uses a global variance metric of the generated parameters as a penalty term in the parameter generation process [18], which significantly improves the synthetic speech quality. Finally, an excitation signal was generated using mixed excitation (pulse plus band-filtered noise components) [19] and PSOLA [20].

At the synthesis stage, for each utterance, context-dependent features were extracted taking into account the target language of each acoustic unit (phoneme). The context-dependent features of each target unit were extracted using the corresponding language-dependent text processing module. Specifically, the context-dependent features of the foreign realisation of an acoustic unit were extracted considering it as a prototypical foreign word with the target context (e.g., trill at the beginning of a word). Finally, the target language of each acoustic unit was added as an additional feature in order to select the appropriate branch at the tree of the bilingual model.

Each bilingual SYNTHETIC stimulus was then synthesised using the bilingual model. The non-accented reference condition consisted of English words produced by the English TTS system. To maximize the naturalness of the synthetic output given that training material consisted of sentences, each stimulus was generated by embedding it in a fixed carrier sentence whose syntax led to natural pauses bracketing the target item. The isolated target item was then extracted using segmentation boundary information made available by the TTS system.

3. Intelligibility and accent judgements

3.1. Methods

Two cohorts of listeners assessed the ability of the three techniques to convey segmental foreign accent and word intelligibility. One group of 9 listeners had English as their L1 while the other group of 21 participants were native Spanish speakers studying English Philology. Each listener heard the set of 108 words in each of the 6 conditions (3 techniques x accent-present/absent). Stimuli were blocked by technique (NATURAL, SPLICED, SYNTHETIC). Within each block accented and non-accented tokens were mixed. The order of the three blocks was counterbalanced across listeners and within each block stimuli were presented in a randomised order. Each block was preceded by a short practice session containing 6 unscored stimuli.

The experiment consisted of two parts. The first part required listeners to type in the word presented. The second part involved making a judgement of degree of foreign accent. Listeners were presented with an orthographic form of each word on the screen and chose an integer value on a 7-point scale labelled "strength of foreign accent" and whose endpoints were labelled "native-like" (1) and "very strong" (7). Listeners were presented with precisely the same set of words in both parts of the experiment. Each part required around 35 minutes to complete. Listeners were paid for their participation.

3.2. Results

Mean word scores and degree of foreign accent (DFA) ratings for the two listener groups are presented in Figure 2. Word scoring took account of homophones but no correction of other errors was undertaken since these may have been attempts by

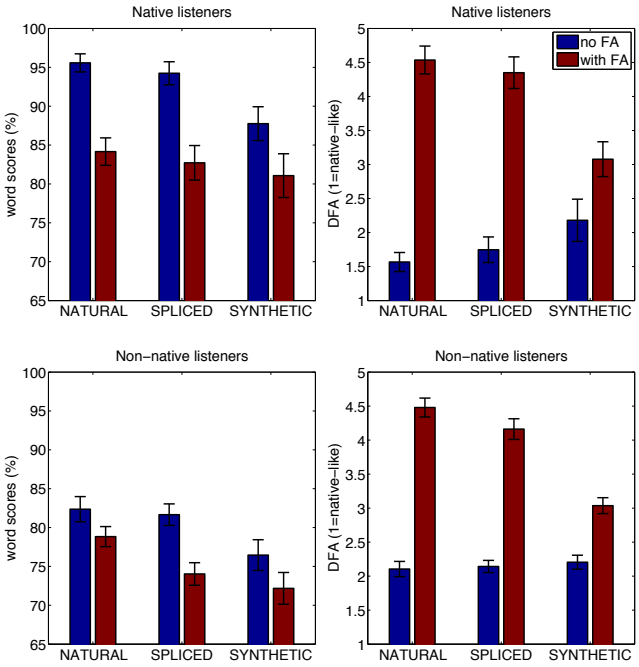

Figure 2: *Intelligibility (left) and foreign accent ratings (right) for the English (top) and Spanish (bottom) listener groups. Error bars indicate* ± 1 *standard error.*

listeners to make sense of accented speech. For statistical analysis, percentages were converted to rationalised arcsine units (RAU) [21]. Due cohort size differences, separate two-factor repeated measures ANOVAs (processing condition $\times$ foreign accent) were carried out for the native and non-native groups, for both RAU scores and mean DFA judgements. We report Fisher's Least Significant Differences (FLSD) to identify statistically-significant differences between factor levels.

For intelligibility, both groups showed a significant interaction between the presence of an accented segment and processing condition [N: $F(2,16) = 7.5, p < .01$, FLSD=4.4; NN: $F(2,40) = 4.6, p < .05$, FLSD=2.2]. For both groups and all conditions the introduction of an accented segment led to a significant drop in intelligibility [$p < .001$ in both cases]. The reduction was somewhat smaller overall for the NN group, suggesting the existence of a (relative) non-native advantage in identifying certain accented words. For the N cohort, accented words were equally-intelligible across processing conditions, while non-accented SYNTHETIC items were significantly less intelligible than non-accented NATURAL or SPLICED words. The NN cohort showed a similar pattern as the N group for non-accented words, with scores that were 11-12 % points lower. Again, for non-accented items the SYNTHETIC versions were less intelligible. Unlike the N group, NN listeners found accented NATURAL items more intelligible than SPLICED or SYNTHETIC accented items.

DFA judgements also revealed substantial interactions between FA and processing condition for both groups [N: $F(2,16) = 112, p < .001, FLSD = 0.22$; NN: $F(2,40) = 130, p < .001, FLSD = 0.14$] and a very clear increase in DFA for the accented words in all conditions [$p < .001$]. Native listeners found NATURAL and SPLICED items to convey a similar degree of foreign accent and SYNTHETIC items rather less so, a similar pattern to that observed for the NN group,

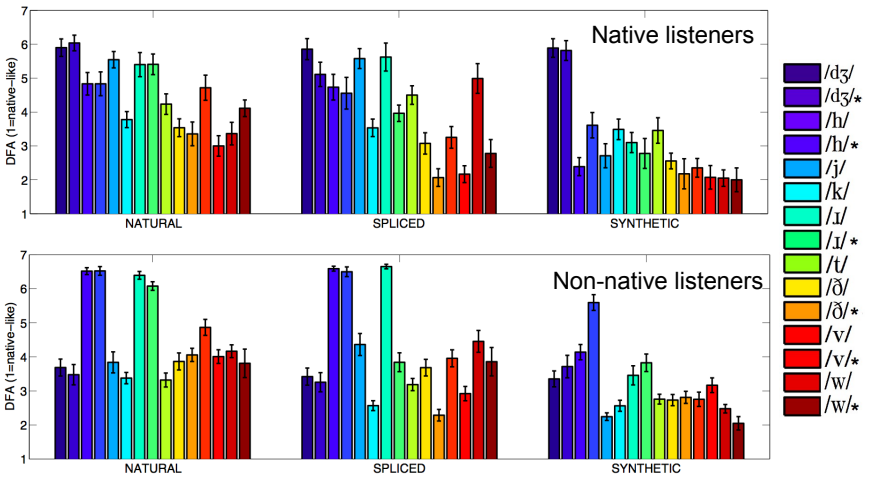

Figure 3: *Accent ratings from native and non-native listeners for accented tokens. Asterisks denote phonemes in medial position.*

although the latter judged the code-switched NATURAL cases to be marginally more accented than the SPLICED tokens. For the NN group, non-accented NATURAL, SPLICED and SYNTHETIC tokens conveyed a similar degree of accent, while the N cohort showed some sensitivity to the different processing conditions, with small but significant decreases in DFA for non-accented NATURAL and SPLICED words re. SYNTHETIC tokens, which were rated similarly by both groups.

The extent to which individual accented phoneme segments affected DFA judgements is shown in Figure 3. Even though the segments chosen were potentially problematic for Spanish speakers of English, we can see that for all three forms of speech wide variation exists in DFA ratings amongst the consonants and that the general pattern of ratings is quite similar across the techniques and listener cohorts. For NATURAL and SPLICED FA, the sound deemed most accented by native listeners was /dʒ/, followed by /j/, /ɹ/ and /h/. For SYNTHETIC FA, /dʒ/ was equally accented, but the others mentioned received considerably lower ratings. For the NN group the distribution is quite different. In all three speech types, /h/ was considered to be the most accented sound, followed closely by /ɹ/.

4. Discussion

Although the pattern of accentedness ratings is similar for NATURAL and SYNTHETIC speech, the latter resulted in lower ratings for many consonants. It may be that the Spanish voice models are not optimal for some L1 sounds in certain positions and substitutions, and that stronger realisations would be found if the voice model made greater use of isolated tokens rather than continuous speech during training. For the NATURAL and SPLICED tokens the speaker was more emphatic (or canonical in the case of the Spanish nonsense words used for splicing) which might explain the higher accent rating for /w/ in SPLICED than NATURAL. In support of this contention one can explain the success of TTS for /dʒ/ and /j/ due to the abundance of /j/ in Spanish sentence material.

For native listeners, the most highly-accented consonant – /dʒ/, realised as a /j/ – is one which results in a phonemic confusion. Note that for non-native listeners this realisation is not considered as accented, probably since it is a feature of their own speech. For the native group, the other possible phonemic confusions resulting from accent, /v, b/ and /ð, d/, rank much lower in terms of accentedness. Indeed, the next two most accented sounds, /ɹ/ and /h/, present Spanish realisations that do not affect meaning. Therefore, although the functional criterion is important for DFA ratings, it is not the only one used by listeners, and a simple 'prototypicality' is also used as a criterion (e.g., the typical Spanish velar/uvular for /h/ and trill for /ɹ/).

For both cohorts, it is notable that both NATURAL and SYNTHETIC tokens produced similar DFA ratings regardless of whether the target consonant occupied an initial or medial position, while the SPLICED condition resulted in more variability for most sounds. It seems likely that the inherent "averaging" quality of HMM-based synthesis is responsible for the consistency here. On the other hand, this over-smoothing property may have limited the ratings of accented tokens. However, it might be possible to use extrapolation techniques [22] to emphasise FA while maintaining the beneficial aspects of robustness in conveying accent in different realisations.

Native listeners found foreign accented speech more intelligible than NN listeners. Nevertheless, on examining intelligibility of accented speech as a proportion of the non-accented forms, we observe a smaller reduction in the face of accent for the non-native cohort. This result is in line with the "matched interlanguage intelligibility benefit" [23]. For accented speech, there are no substantial differences between the two listener groups, which supports our earlier findings [24]. The difference in sensitivity to accent stems from the native cohort's greater discrimination between non-accented speech in the three styles.

5. Conclusions

The current study demonstrated several generative approaches to localised speech alterations which result in a reasonably convincing degree of foreign accent for consonants. Future work will extend the approach to vowels and explore ways to enhance accentedness within the HTS framework. One practical application of the methods is in the design of tools aimed at improving L2 phonological acquisition in language learners.

Acknowledgements. MLGL, RPR & MC were funded by the Spanish MINECO project DIACEX (FFI2012-31597). RBC was partially funded by EU grant 287678 and INAPRA (MICINN, DPI2010-21247-C02-02).

6. References

[1] L. D. Hahn, "Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals," *TESOL Quarterly*, vol. 38, pp. 201–223, 2004.

[2] B. Zielinski, "The listener: No longer the silent partner in reduced intelligibility," *System*, vol. 36, pp. 69–84, 2010.

[3] K. Takima, R. Port, and J. Dalby, "Effects of temporal correction on intelligibility of foreign-accented English," *J. Phonetics*, vol. 25, pp. 1–24, 1997.

[4] M. Munro and T. M. Derwing, "Modeling perceptions of the accentedness and comprehensibility of L2 speech," *Studies in Second Language Acquisition*, vol. 23, pp. 451–468, 2001.

[5] O. Kang, D. Rubin, and L. Pickering, "Suprasegmental measures of accentedness and judgements of oral language proficiency in oral English," *The Modern Language Journal*, vol. 94, pp. 554–566, 2010.

[6] M. Munro, T. Derwing, and C. Burgess, "Detection of NN speaker status from content-masked speech," *Speech Communication*, vol. 52, pp. 626–637, 2010.

[7] T. Derwing and M. Munro, "Accent, intelligibility, and comprehensibility: Evidence from four L1s," *Studies in Second Language Acquisition*, vol. 19, pp. 1–16, 1997.

[8] M. Munro and T. Derwing, "The functional load principle in ESL pronunciation instruction: An exploratory study," *System*, vol. 34, pp. 520–531, 2006.

[9] F. Casacuberta, R. Garca, J. LListerri, C. Nadeu, J. M. Pardo, and A. J. Rubio, "Desarrollo de corpus para investigacin en tecnologa del habla (ALBAYZIN)," *Procesamiento de Lenguaje Natural*, vol. 12, pp. 35–42, 1992.

[10] R. A. Clark, K. Richmond, and S. King, "Festival 2 – build your own general purpose unit selection speech synthesiser," in *Proc. 5th ISCA Workshop on Speech Synthesis*, 2004.

[11] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. International*, vol. 5, no. 9/10, pp. 341–345, 2002.

[12] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, *The HMM-based speech synthesis system (HTS) Version 2.1*, 2008, http://hts.sp.nitech.ac.jp/.

[13] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 1, pp. 66–83, 1 2009.

[14] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication*, vol. 52, no. 5, pp. 394–404, 2010.

[15] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[16] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[17] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorigthms for HMM-based speech synthesis," in *Proceedings ICASSP 2000*, Jun. 2000, pp. 1315–1318.

[18] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[19] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *2nd MAVEBA*, Sep. 2001.

[20] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–468, 1990.

[21] G. Studebaker, "A rationalized arcsine transform," *J. Speech Hear. Res.*, vol. 28, pp. 455–462, 1985.

[22] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[23] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *J. Acoust. Soc. Am.*, vol. 114, pp. 1600–1610, 2003.

[24] F. Gallardo, M. L. Garcia Lecumberri, and E. Gomez, "The assessment of FA and its communicative effects by naive native vs. experienced non-native judges," *International Journal of Applied Linguistics*, 2014.