Voice morphing and the manipulation of intra-speaker and cross-speaker phonetic variation to create foreign accent continua: A perceptual study

John Ingram [1], *Hansjörg Mixdorff*[2] *and Nahyun Kwon* [1]

[1] School of EMSAH, University of Queensland, Brisbane, Australia
[2]Department of Informatics and Media, BHT University of Applied Sciences, Berlin, Germany
j.ingram@uq.edu.au, mixdorff@beuth-hochschule.de

Abstract

The STRAIGHT system of voice morphing was used to create voice continua of (Korean) accented Australian English, intended to simulate phonetic variation ranging from 'heavily accented' to 'unaccented' (native-like) Australian English, employing dimensions of intra-speaker and cross-speaker variation to yield a range of synthetic voices. These synthetic voices were evaluated against actual samples of Korean accented English, both re-synthesized and non-re-synthesized, in a series of three perceptual rating experiments by native listeners of Australian English. The questions of central interest in this preliminary investigation are: (a) the method of creating the phonetic continua and the respective roles of intra- versus cross-speaker variability in simulating degrees of foreign accent, (b) the success of the STRAIGHT method for creating hybrid voices, compared with 'natural' tokens of accented utterances, and (c) the impact of the re-synthesis method (required for voice morphing) upon perceptual ratings of foreign accent by native listeners. The ultimate objective of this research is to assess the impact of segmental and prosodic features on the perception of foreign accent and intelligibility of L2 learners' speech, where the source (Korean) and target (English) languages pose significant difficulties of segmental and prosodic transfer.

1. Introduction

There appears to be considerable individual variation in peoples' ability to acquire native-like pronunciation in a second language when it is acquired in linguistic maturity. This is particularly noticeable where the source and target languages differ typologically in prosody as well as segmental phonology. Accent is a strong indexical marker of speaker identity. It is a moot point as to how clearly attributes of personal voice quality, by which we identify speakers are separable from those phonetic features which mark their linguistic affiliation (accent) and which might be captured in a narrow phonetic transcription of their speech. More generally, the role of individual speaker characteristics in speech recognition and whether there is need of a signal-conditioning stage of 'speaker normalization' in speech recognition has been a lively source of debate for theories of speech perception in recent years [1].

When we sought to use Voice Morphing technology to create phonetic continua of accented English, we were confronted with a decision as to whether to attempt to isolate phonetic variation that might be reflected in a speaker-neutral phonetic transcription, from the 'non-phonetic' variation that marks individual speaker or voice identity. The most straightforward method of creating an accent continuum was to morph across two speakers' voices, representing the 'foreign' and the 'native' ends of the accent continuum, without any attempt to control variation in personal voice characteristics, beyond matching the gender of the two speakers forming the 'accented' and 'unaccented' ends of the voice morphing continuum. We refer to this as 'cross-speaker' morphing.

To generate an 'intra-speaker' accent continuum, controlling for personal voice characteristics, we required a high level bilingual speaker, capable of pronouncing the target utterances with native-English-like fluency and with as little trace of 'foreign accent' as possible, but also with the ability to read phonetically matched nonce sentences, presented in the source language orthography, with native-like fluency. For example, each English target sentence was assigned a Korean transliteration, which our bilingual speaker was instructed to read fluently with a 'Korean accent', as if the nonce utterance were a meaningful Korean sentence. By analysing the phonetic sequence of the English target sentence in terms of the phonological system of Korean, we made sure that we only employed legal phonemes and syllables of that language. In our mind this is the most serious case of L1 interference one can possibly imagine, even if in reality and for a given speaker not all errors actually occur. However, since it is a controllable condition, we regard it as a better starting point than simply referring to an effervescent snap shot of what a foreign learner of English will produce at a given moment:

English target: A mask covered the soldier's face and mouth
Korean transliteration: 어 마스크 커버드 더 쏠저스 페이스 앤드 마우스
Phonetic transcription: [ma.s .k k .b .d t s'ol.c .s p e.i.s n.t ma.u.s]

The pronunciation of the transliteration was intended to simulate – how plausibly remains to be determined – the phonetic characteristics of a beginning level Korean learner of English. The contrasting pronunciations of the English target and the Korean transliteration were then deployed as end points for constructing a synthetic Korean – English accent continuum. Both sets of cross-speaker and intra-speaker morphed utterances were to be embedded with natural accented English tokens elicited from (Korean) L2 speakers of English in a listening experiment using native English listener judgments of foreign accent strength with the aims (as stated above) of evaluating the methods of creating accent variability, the quality of the Voice Morphing technique, and the impact of re-synthesis on the listener ratings of accent strength and intelligibility.

Hence we performed two experiments: One employing only natural foreign accented stimuli (Experiment 1) for testing inter-rater reliability and one employing the morphed stimuli

(experiments 2 and 3) embedded within a selection of natural sentences

2. Speech Material and Method of Manipulation

Stimuli for experiments 2 and 3 were generated by applying Tandem-STRAIGHT-based morphing from three types of stimuli:

(1) Korean Speaker $CWK_{Kor.}$, Korean transliterations
(2) Korean Speaker $CWK_{Eng.}$, English sentences
(3) Australian English Speaker $JI_{Eng.}$, English sentences

The morphing procedure requires temporal reference points serving as anchors. The anchors were produced by manually segmenting the utterances on the phoneme level and supplying the mid point of each segment as an additional anchor. Since, however, the source as well as the target utterance may contain different segments, these still have to be marked (with durations close to 0) in the other utterance in order to create congruent sequences of morphing anchors. It is important that the locations of these additional "ghost" anchors, which due to a restriction in the morphing algorithm cannot have zero time spacing, are selected carefully, optimally during instances of pauses or during adjacent sounds that have the same type of excitation (voiced/unvoiced) as the respective segment in the other utterance. Otherwise noticeable artefacts can arise when source and target are mixed.

We generated morphing sequences at 0, 33, 67 and 99% ratios between (1) and (2), as well as (1) and (3). In addition we created a morphing sequence between (1) and (3) in which only the prosodic features (F0 contour and timing) of JI were combined with the spectral features of JWK at ratios of 0, 33, 67 and 99%.

Due to some inevitable effects of the morphing on the acoustic quality of the signals we decided to band-limit the stimuli to 300-4000 Hz and add some dizzering white noise of -40 dB SNR. As a reference to the morphed stimuli we included a number of natural accented utterances which where STRAIGHT analysed, re-synthesized and subjected to the same band-pass filtering and dizzering.

3. Experimental Designs

3.1. Design Experiment 1

The first experiment was designed to evaluate the anchor point stimuli that were to be used to construct the intra-speaker and cross-speaker accent morphing continua. The aim was to locate these reference stimuli in relation to natural Korean-accented English tokens. Six target utterances were selected by NK (third author, this paper) from a larger set of utterances used in a previous study (Ingram and Nguyen, 2007) The six sentences selected for their likelihood of eliciting Korean transfer effects were:

1. *A mask covered the soldiers face and mouth.*
2. *The queen was sleeping in the royal tent.*
3. *The world's driest continent is Australia.*
4. *They hung blue-bells from the eves of the greenhouse.*
5. *They used to hunt elephants for their tusks and hides*.
6. *They wanted to migrate to a friendly society*.

Speakers: Two fluent Korean-English bilingual speakers were recruited to produce Korean and English anchor stimuli. CWK is a Korean-born male university lecturer, in a Korean language program, aged 45 years and has been continuously resident in Australia for the past 20 years. He has native-like fluency in English, with a mild but detectable Korean accent. NK is a Korean-born female, aged 23 years, with two years residence in Australia. She is an MA student in Linguistics and third author of this paper. She speaks quite fluent English with an unmistakable Korean accent. Both speakers have extensive phonetics training and a critical appreciation of language transfer effects. CWK has extensive experience in multi-media production for Korean language teaching. Each of the two Korean speakers was matched with the voice of a native speaker of Australian English. JI, first author this paper was paired with CWK and LW, a female post-graduate student of linguistics and native speaker of Australian English, provided an English voice match for NK. Four Korean learners of English, two males and two females, all 'overseas students' enrolled as undergraduate students at Griffith University, with less than two years residence in Australia, but of varying English fluency and experience were recruited to provide Korean – English accented productions of the six target utterances.

Sentence elicitation: For production of the English anchoring stimuli, CWK, NK and LW were asked to listen to JI's productions of the target sentences and to produce their own, at roughly the same pace but in their natural (English) voice. Multiple versions of the target sentences were elicited and JI selected the most natural and English-sounding token. There was little variation among the token productions for CWK but more variation in the case of NK.

For production of the Korean anchoring stimuli (by CWK and NK), which were elicited after they had produced the English anchor set, the two Korean speakers were asked to read the transliteration target sentences fluently, as though they represented sensible Korean sentences. This proved to be not necessarily a straightforward task. There was inevitably some interference produced by familiarity with the near-homophonous English counterparts, which they had just previously practiced. In fact, in order to produce a fluent reading of the Korean nonce string, it may have been necessary to retain some trace in short term memory of the prosodic contour of the English counterpart. However, both Korean readers succeeded in producing fairly fluent readings of the nonce Korean utterances.

For production of the target sentences by the four Korean learners of English an elicitation strategy that had been successfully employed in previous experiments [2] was adopted, which was intended to deflect subjects' attention from any 'deficiencies' in their English pronunciation and encourage them to employ their natural (English) speaking voice. Instead of presenting the target sentences directly to be read, subjects were given a syntactic paraphrase of the target and asked to produce a paraphrase cued by the first word:

Example of the paraphrase task:
The soldier's face and mouth was covered by a mask.
A mask _____

The paraphrase task focuses the speaker's attention on the linguistic and not the pronunciation aspects of the task, helping to ensure that they use a more habitual unmonitored speaking style. The ten 'voices' used in the production of anchor stimuli for voice morphing (voices 1-6) and the 'naturally accented' Korean-English speakers (voices 7-10) against which they would be evaluated, are shown in Table 1, together with predictions made by the experimenters as to how strongly accented each voice would be rated by Australian English listeners. We refer to these voices used in experiment 1 as the 10 'non-morphed' voices.

Table 1: *Experiment 1, voices and predictions, obtained accent ratings (right-most column, μ/σ, mapped to seven-point scale).*

voice	classification	predictions	rating
1	CWK_{Eng}(Engl.imitations)	mild accent	3.1/1.1
2	CWK_{Kor} (transliterations)	strong accent	5.4/1.3
3	$NK_{Eng.}$ (Engl. imitations)	mild- moderate	3.3/1.2
4	$NK_{Kor.}$(transliterations)	strong accent	5.8/1.2
5	$LW_{Eng.}$(Native Aust.Eng.)	no accent	1.2/.6
6	$JI_{Eng.}$ (Native Aust. Eng.)	no accent	1.0/.2
7	HS –Korean-Engl. female	strong accent	5.2/1.2
8	HB –Korean Engl. female	mild accent	2.5/1.3
9	BE – Korean Engl. male	mild accent	3.9/1.3
10	CM – Korean Engl. male	mild accent	4.0/1.4

Accent rating experiment: Native Australian English listeners were recruited from a large introductory linguistics class at the University of Queensland, and given course credit for participation in a short experiment, run over the web, that involved their listening to 30 spoken sentences, which they were to rate for strength of 'foreign accent' on a five-point scale and make some observations about difficulty of word comprehension. Participants were provided written equivalents of the utterances to be judged. There were 60 sentences to be rated (six target sentences x ten speaking voices). Because of the need to keep the experiment short (15-20 minutes), the full set of items had to be distributed over two listener groups. Items were distributed across listener groups such that every listener heard multiple tokens of every sentence, but only tokens from half of the speakers (voices).

3.2. Design Experiments 2 and 3

Two additional sets of experimental voices were constructed via the STRAIGHT morphing system and these were rated by listeners drawn from the same subject pool as experiment 1, native listeners of Australian English. Hence, 250 students were randomly allocated to one of six (3x2) listening groups in approximately equal numbers. To keep the number of stimuli to a manageable size, only 4 steps were used to define points on an accent morphing continuum: (0, 33, 67 and 99%). The male Korean speaker (CWK, paired with JI's voice for cross-speaker morphing) yielded cleaner stimuli for perceptual evaluation than the female Korean and Australian speakers (NK and LW). Consequently, only male morphed voices were evaluated in experiments 2 and 3.

A seven-point Likert scale of accent strength was used for rating the morphed utterances. We chose a seven point scale instead of five points as in the first experiment because we expected a narrower perceptual spacing for the morphed stimuli. The number ratings on which each accent rating is based (N) is large and varies because of the composition of the stimulus sets and the fact that ratings for morphed stimuli are aggregated over both morphing experiments. Largish groups of listeners were used in an effort to ensure that robust and discriminating accent scores would be obtained from potentially noisy data. In the results that follow we report ratings of degree of foreign accent averaged across the six target sentences and all listeners.

4. Results and analysis:

4.1. Results Experiment 1

We performed split-correlation reliability analysis by dividing the utterance-wise judgments into two participant groups of equal size, yielding a cross-correlation between the two groups of .978 ($p<.001$). This shows that the ratings are stable. Mean accent ratings for the ten 'voices' averaged across all sentences and listeners are shown in the right-most column of Table 1 giving means and standard deviations. For the sake of better comparison with experiments 2 and 3 the results yielded on the five-point scale were mapped to a seven-point scale.

These accent rating scores apply to the 'un-morphed' stimuli used to construct end points for the intra-speaker and cross-speaker accent continua. Comparisons within this stimulus set enable one to evaluate where the transliteration productions fall on the continuum of accent ratings in comparison with the pattern of predicted accent ratings on the left side of Table 1. As can be seen, the order of numerical values matches the predictions well, though mild 'accents' range between scores of 2 and 3.

4.2. Results: Experiments 2 and 3.

The results of primary interest concern the efficacy of accent morphing under three morphing conditions: a) cross-speaker morphing (Table 2) using the full set of acoustic parameters, b) Intra-speaker morphing (Table 3) using all parameters, and c) Cross-speaker morphing using only the prosodic parameters of timing and f_0 contour (rhythm and voice pitch) (Table 4).

Table 2*: Cross-speaker morphing: all features Korean – English (CWK_{Kor} - JI_{Eng})*

morph. ratio [%]	Accent rating		
	Mean	N	s.d.
0	4.71	347	1.29
33	4.23	457	1.26
67	2.29	365	1.28
99	1.59	408	1.01

Table 3: *Within-speaker morphing: all features, Korean – English (CWKKor - CWKEng).*

morph. ratio [%]	Accent rating		
	Mean	N	s.d.
0	5.07	212	1.25
33	4.41	370	1.20

67	3.12	559	1.34
99	3.19	394	1.33

Table 4: *Cross-speaker morphing: prosodic only Korean – English (CWK_{Kor} - JI_{Eng})*

morph. ratio [%]	Accent rating		
	Mean	N	s.d.
00	4.71	347	1.21
33	4.23	347	1.21
67	4.12	347	1.25
99	4.02	344	1.27

It is evident from the descriptive statistics reported in Table 2 – 4 that Cross-speaker morphing using the full set of acoustic parameters is the most effective treatment for changing listeners' mean rating scores across the four morphing ratios that were tested, changing the ratings from 'moderate – fairly strong' foreign accent to 'no detectable – slight foreign accent', in line with predictions from the un-morphed anchor point stimuli. This result may be seen as a vindication of the STRAIGHT morphing method. The spacing between the judgments at different morphing levels, however, is not even and we witness the largest gap between 33 and 67%.

By contrast, the within-speaker morphing treatment changed listeners' accent ratings from 'fairly strong accent' (for CWK's productions of the transliteration productions) to 'mild foreign accent' – also in line with predictions and ratings from the un-morphed anchor stimuli (Table 1). Against our expectations, however, the case 67% yields slightly better ratings than the case 99% which is supposed to be equal to the 'English' version. However, this result is not statistically significant and reversed for some of the sentences.

The accent rating shift obtained under voice morphing (and with telephone band filtering) is comparable with the accent differences observed for the un-morphed (and not band limited) anchors of CWK's transliteration productions and his English voice pronunciation. This result also speaks to the efficacy of the accent morphing technique.

Cross-speaker morphing using only prosodic parameters (timing and f0 contours, Table 4) was clearly less efficacious than full parameter manipulation, changing listeners' perception of foreign accent only marginally – but in a coherently graded fashion across the four morphing ratios. This result is an important indication of the contribution of prosody to the percept of foreign accent. Whereas with the full set of features (Table 2) a range of 3.12 on the Likert scale is covered, the figure is .69 for prosody only, hence a non-negligible contribution of 22.1%.

5. Discussion and Conclusions

Space does not permit presentation of an inferential statistical analysis of the results, which will be available for the oral presentation. However from the descriptive statistics presented above, we can conclude that the STRAIGHT voice morphing technique succeeded in creating plausible accent continua constructed from both cross-speaker and intra-speaker phonetic variation. The morphing ratio, however, can only serve as a coarse indicator of the effects the procedure will have on the ultimate accent rating. In the case of speaker CWK we found, for instance, that at a ratio of 99% which was supposed to provide the least accented version, stimuli were perceived similarly as at 67%. In addition there seems to be a large perceptual gap between 33 and 67%. Auditory comparison of 67 and 99% stimuli in the 'wrong order' suggests that the intonation contour is more expressive in 99% (actually closer to the original) and therefore less acceptable than the flatter contours in 67%. Both these utterance types, however, are close to the mean rating of 3.1 found for CWK in experiment 1 (Table 1, line 2).

Another result that needs yet to be explained is the fact that at 4.71 the 0% stimulus in the CWK-JI series is judged less foreign accented than the 0% stimulus in the within-CWK series at 5.07. If a ratio of 0% denotes that the source utterance (which is the same in both cases) contributes a 100% to the mixture, both should be rated similarly.

From our preliminary findings it appears that cross-speaker phonetic variation is much more effective than manipulation of intra-speaker phonetic variation for creating wide continua of perceived foreign accent. Extrapolating from these findings, it may well be the case that even for highly fluent bilinguals, accent continua ranging from 'strongly accented' to 'native like' or 'un-accented' pronunciation will be very difficult to achieve. This may have implications also for the (in)separability of personal and linguistic-phonetic features in speech and for the architecture of models of speech perception.

Finally, there are preliminary indications that prosody may play a limited role in determining perceptions of foreign accent. Our method, however, facilitates the quantification of various contributing factors to the percept of foreign accent. But more detailed work is needed on this point and on the problems that arise from attempts to analytically manipulate prosodic and segmental features in voice synthesis for foreign accent modeling.

In future work we plan to examine stimuli with temporally variable morphing ratios, as well as other native languages than Korean. Employing variable morphing ratios will enable us to consistently simulate and examine the effect of certain accent phenomena such as vowel or consonant replacements. We will also continue to seek to recruit truly bi-lingual subjects capable of producing the two languages involved without perceivable accent (assuming such persons exist).

6. Acknowledgements

We would like to thank Hideki Kawahara for supplying the STRAIGHT analysis/re-synthesis and morphing tools and his generous and patient support of this work.

7. References

[1] Johnson, K. & Mullennix, J. W. (Eds.) *Talker variability in Speech Processing*. San Diego: Academic Press, 1997.

[2] Ingram J., & Nguyen T., "Vietnamese accented English: Foreign Accent and Intelligibility judgement by listeners of different language backgrounds." Proceedings of *"TESOL in the internationalization of higher education in Vietnam" Conference*. Hanoi, Vietnam, 2007.

[3] Kawahara, H., "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown." Proceedings of ICASSP 2009, Taipei, Taiwan.