The Role of Prosody in the Perception of US Native English Accents†

Ayako Ikeno and John H.L. Hansen

The Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas; Richardson, Texas 75083, USA,
{Ayako.Ikeno,John.Hansen}@utdallas.edu http://crss.utdallas.edu

ABSTRACT

A wide range of aspects are contained within the speech signal which provides information concerning a particular speaker's characteristics. Accent is a linguistic trait of speaker identity. It indicates the speaker's language and social background. The goal of this study is to provide perceptual assessment of accent variation in US native English. The main issue considered is how different components of prosody affect accent perception. This perceptual study employed an ASHA certified acoustic sound booth using 73 listeners (53 male, 20 female). The results from these perceptual experiments indicate the importance of prosody in combination with availability of utterance content via speech signal or transcripts. The trends also indicate that listeners' decisions are influenced by conceptual representation of prototypical accent characteristics, such as "people from New York talk fast." These observations suggest that listeners use both bottom-up processing, based on the acoustic input, and top-town processing, based on their conceptual representation of prototypical accent characteristics. Those processes are multi-dimensional in that listeners use utterance content (e.g., meaning or comprehensibility) as well as accent characteristics in the acoustic input even though our experiment focuses on pronunciation features and does not include word selections that are dialect dependent. These findings contribute to a deeper understanding of the cognitive aspects of accent variation, and its applications for speech technology, such as accent classification for speaker identification or speech recognition.

Index Terms: accent, dialect, prosody, perception.

1. Introduction

Accent (or dialect) is a crucial factor for speech technology in various areas including business, security, and language education, as illustrated in Fig. 1.

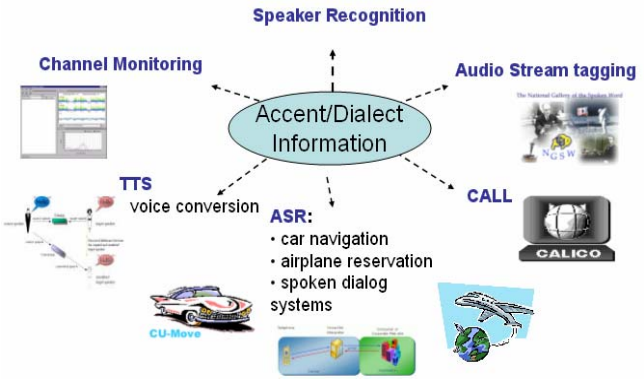

Figure 1. Applications that can benefit from Automatic Recognition of Accent and Dialect Information

For example, identification or classification of speaker accent can provide useful information for Automatic Speech Recognition and Understanding. Accent and dialect characteristics can also provide important information concerning speaker identity. Automatic Accent Identification therefore, is an important part of technological application of accent characteristics for forensics and security as well as communication. Investigating the cognitive aspect of accent variation is important, since factors based on how humans categorize accents provide meaningful insight and knowledge for further development of accent classification algorithms [1, 2, 3] and speech technology.

The goal of this study is to identify speech characteristics that distinguish different accents perceptually across a variety of US native English accents. Specifically, this study will examine how different components of prosody affect accent perception.

Accent and *dialect* both refer to linguistic variation of a language. Use of these terms can be ambiguous. In this paper, we use the term *accent* as defined in Crystal[4] – "The cumulative auditory effect of those features of *pronunciation* which identify where a person is from regionally and socially. The linguistic literature emphasizes that the term refers to *pronunciation only*, is thus distinct from dialect, which refers to grammar and vocabulary as well."

Previous studies on perceptual classification of US native English accents have shown that listeners are able to determine differences between southern and non-southern accents of American English when natural speech is provided (e.g., [5, 6]). Clopper[5] also reported that southern speakers had significantly longer vowel duration compared to New England speakers. This suggests that prosody, duration or tempo in this case, may contribute to accent perception. In addition, it has been suggested that, in the case of foreign accent detection (e.g., US native English vs. German accented English), f0 contour provides cues even with low-pass filtered, unintelligible speech (e.g., [7, 8, 9]).

The analysis in this paper focuses on US native English accent variation and addresses the following two issues: 1) how accurately are listeners able to perceive the variability of accents with only prosodic information (e.g., using low-pass filtered speech without text content), and 2) how prosody, in particular f0 contour or tempo, contributes to the perception of accent. We also discuss how bottom-up processing (e.g., acoustic signals) and top-down processing (e.g., prototypical characteristics of accent) influence accent perception, and the complexity involved in this process.

2. Data, Listeners & Experiments

The test data represent the following native English accents – California, (US west), Mississippi (US south), New York (US northeast). The data set is composed of approximately 3 second

† This work was supported by the U.S. Air Force Research Laboratory, Rome NY under contract No. FA8750-05-C-0029, and RADC under contract No. FA8750-05-C-0029, and by University of Texas at Dallas under Project Emmitt.

long sentences extracted from spontaneous speech produced by male speakers in CALLFRIENDS and CALLHOME (available from LDC, www.ldc.upenn.edu). Since this study focuses on *accent* (or pronunciation) characteristics, the speech data did not include words that are dialect dependent, which may provide additional cues for perceptual classification tasks.

The number of listeners totals 73 (53 males, 20 females). The listeners were all undergraduate students at Univ. of Colorado at Boulder with ages ranging from 18 to 23 years old. A preliminary pure-tone hearing screening was preformed for each subject at ASHA suggested frequencies of (500, 1000, 2000, 4000, 8000)Hz., and all participating listeners passed with normal hearing.

The listening test was conducted individually in an ASHA certified double-wall sound booth using an interactive computer interface and a Bose QC2 headset. Tasks were perceptual classification of US native English accent using natural and processed speech samples. *Task 1* focuses on the effect of the presence of the original f0 contour when either the original phonetic content (i.e., vowel and consonant characteristics) or suppressed phonetic content (via low-pass filtering (LPF)) is included in the speech samples. *Task 2* focuses on the effect of the presence of the original tempo and f0 contour when the original phonetic content is present.

3. Accent Perception Task 1 Results

For Task 1, the listeners were asked to classify accents using speech samples with the following 4 Conditions:

1) ***Type 1a (Filter & Monotone)*** – original duration and energy content, with processed flat f0 contour (e.g., normalized f0 contour resulting in monotone speech), and suppressed phonetic content (i.e., LPF to remove text content),
2) ***Type 1b (Filter)*** – original duration, energy and f0 contour, with suppressed phonetic content using LPF,
3) ***Type 2a (Monotone)*** – original duration, energy, and phonetic content, with processed/normalized flat f0 contour,
4) ***Type 2b (Natural)*** - original, unprocessed speech.

A total of 36 speech samples, 12 per accent type, were presented for each condition. Phonetic content of the speech was suppressed by low-pass filtering (LPF) at 225Hz for Conditions 1 and 2. Monotone speech (flat f0 contour at mean f0 value of each speech sample) was produced using Time-Domain PSOLA [10, 11].

Listener Group 1 (25 listeners) was provided with only audio samples, and ***Listener Group 2*** (24 listeners) was provided with both audio samples and text transcripts for low-pass-filtered, unintelligible speech with suppressed phonetic content (Type 1a - Filter & Monotone, and Type 1b - Filter).

Listeners were also asked to indicate a confidence rating on a 1-5 point scale for each selection. Listener confidence was rated as shown in Fig. 2.

1	2	3	4	5
not sure at all		somewhat sure		Absolutely sure

Figure 2. Confidence Ratings for Accent Listener Testing

The result in Fig. 3 illustrates the effect of the presence of the original f0 contour on classification accuracy, using Type 1 (low-pass filtered, "unintelligible" with text content removed) speech. When listeners have no access to transcripts (Listener Group 1), the presence of the original f0 contour contributes to classification accuracy 7% relative on average. When transcripts are available (Listener Group 2), the presence of f0 contour improves the classification accuracy 18% relative on average. This indicates that f0 contour provides more meaningful information when listeners know the content of speech (e.g., utterance meaning, word sequence, and possible sequence of phonemes, etc.) although the degree of its effect may vary depending on accent type.

An analysis of confusability between two of the three accents in this study (e.g., CA is misperceived as MS) also indicates that the presence of the original f0 contour impacts the accuracy of perceptual accent classification when transcripts are available (Listener Group 2, a 23% relative gain on average).

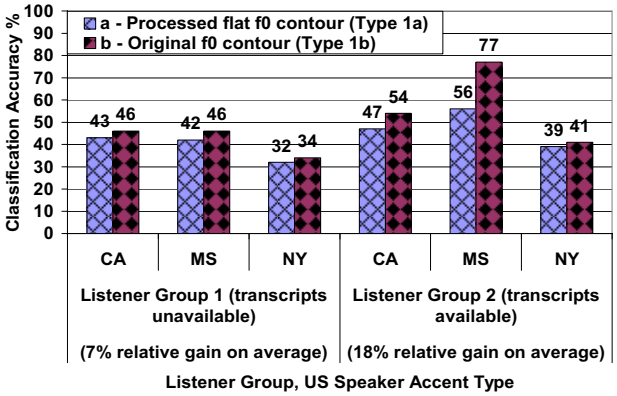

Figure 3. The Effect of the Presence of the Original F0 Contour on Perceptual Accent Classification Accuracy with Unintelligible (Type 1) Speech. The clearest impact is observed in the case of Mississippi (MS) accent with Listener Group 2. The accuracy improves from 56% (with the processed flat f0 contour) to 77% (with the original f0 contour).

The effect of transcript availability is more clearly illustrated in Fig. 4. For both Type 1a (processed flat f0 contour) and Type 1b (original f0 contour) speech, the availability of transcripts influences the accuracy of perceptual accent classification. With Type 1a speech (Filter & Monotone), the accuracy improves 17% relative on average. With Type 1b speech (Filter), the result shows a 23% relative gain on average. This trend further indicates the importance of having access to speech content, such as meaning, word and phoneme sequence, in accent perception.

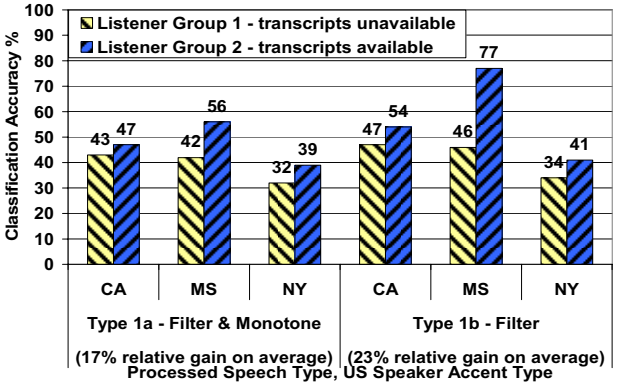

Figure 4. The Effect of Availability of Transcripts on Perceptual Accent Classification Accuracy. The clearest impact can be seen in the case of Mississippi (MS) accent using speech with the original f0 contour. The accuracy improves from 46% (without transcripts) to 77% (with transcripts).

The confusability between two of the three accents (e.g., NY is misperceived as CA) decreases significantly with access to transcripts, especially when the original f0 contour is included in the speech (Type 1a – 12% relative gain on average, Type 1b – 29% relative gain on average). These observations further suggest that the presence of the original f0 contour contribute to accent perception even when details of vowel and consonant characteristics are not present in the speech, although the accuracy may not reflect this effect directly. This understanding is important since it implies that the f0 contour can provide meaningful cues for accent classification even when

the speech signal is not clean due to environmental noise or channel issues.

Another important trend here is that, when transcripts were available, listeners were able to differentiate southern accent (MS) from non-southern accent (CA and NY) 77% of the time with only prosodic information (i.e., energy, duration, and f0 contour). This accuracy is even higher than the accuracy for California accent and New York accent with original (Type 2b - Natural) speech (up to 75% and 63%, as illustrated in Fig. 5). This indicates that characteristics of Mississippi accent may be prosodically distinct compared to non-southern accents.

The result with Type 2a (Monotone) and Type 2b (Natural) intelligible speech shows a similar but stronger trend. Regardless of the presence or absence of the original f0 contour, Mississippi accent is accurately classified 97% to 98% of the time, as illustrated in Fig. 5. This result supports Clopper's[5] observation that listeners can differentiate southern accent vs. non-southern accent (when unprocessed, natural speech is presented). Our result also confirms the importance of phonetic content (vowel and consonant characteristics) in accent perception (e.g., [5, 6]).

On the other hand, even when the original phonetic content is included in the speech, the presence of the original f0 contour contributes to accent classification (6% relative gain on average).

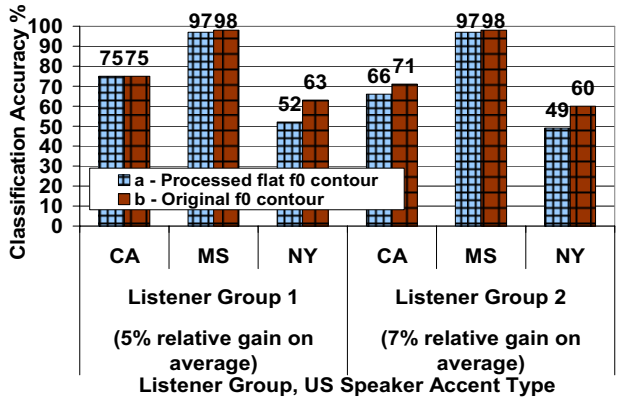

Figure 5. The Effect of the Presence of the Original f0 Contour on Perceptual Accent with Intelligible (Type 2) Speech. For example, MS accent was correctly classified 97% of the time with Type 2a (Monotone) speech, and 98% o the time with Type 2b (Natural) speech.

The effect of the presence of the original f0 contour with Type 2 intelligible speech is marginal, as can be seen in Fig. 5. However, the analysis on confusability, similarly to the case of Type 1 unintelligible (LPF text content removed) speech, shows a large reduction with the presence of the original f0 contour. The confusability between two of the three accents (e.g., CA accent is misperceived as MS accent) decreases 33% relative on average when the original f0 contour is included, although the accuracy does not directly reflect this improvement. Furthermore, to support our observations on classification accuracy and confusability, confidence ratings show similar trends. Although the numbers are not directly comparable to the accuracy or confusability measures, confidence ratings are higher when the original f0 contour is present in the speech. Overall, these trends together indicate that the presence of the original f0 contour contributes to accent perception. These findings suggest the importance of incorporating f0 characteristics for automatic accent classification and identification.

4. Accent Perception Task 2 Results

The analysis of Task 2 results focuses on the effect of the presence of the original tempo and f0 contour. For this task, the listeners were asked to classify accent types using the speech samples with the following 5 conditions:

1) ***Type 1 (Fast & Monotone)*** - processed fast tempo and flat f0 contour
2) ***Type 2 (Slow & Monotone)*** - processed slow tempo and flat f0 contour,
3) ***Type 3 (Fast)*** - processed fast with the original f0 contour,
4) ***Type 4 (Slow)*** - processed slow with original f0 contour,
5) ***Type 5 (Natural)*** - original, unprocessed speech.

Both tempo and f0 contour were adjusted using TD-PSOLA.

The classification accuracy from Task 2 result shows a 5% relative gain on average as illustrated in Fig. 6. The confusability between two of the three accents also shows a 17% relative reduction on average. Although the accuracy does not directly reflect a reduction in confusability, these trends suggest that tempo together with the f0 contour influence accent perception. Although the classification performance is influenced differently depending on the accent type, there are consistent trends observed through accuracy and confusability rates. That is, listeners' decisions were influenced by conceptual representation of prototypical accent characteristics, which may or may not be authentic. For example, when Type 1 (Fast & Monotone) speech or Type 3 (Fast) speech was presented, New York accent was accurately perceived as New York accent more often (53%) compared to Type 2 (Slow & Monotone) speech and Type 4 (Slow) speech (53%). In the cases of confusability, New York accent was misperceived as Mississippi Accent less often (4% and 2%) compared to when Type 2 (Slow & Monotone) speech or Type 4 (Slow) speech was presented (11% and 6%). In addition, when Type 1 (Fast & Monotone) speech was presented, California accent was misperceived as New York accent more often (32%) compared to Type 2 (Slow & Monotone) speech (25%). It was also the case that California accent was more often misperceived as Mississippi accent (5%) with Type 2 (Slow & Monotone) than with Type 1 (Fast & Monotone) speech (2%). Listeners might have had a prior conceptual representation that people from New York speak fast and people from Mississippi speak slowly. These results suggest that tempo (alone or together with f0 contour) contributes to accent perception, and that conceptual representations of prototypical accent characteristics influence listeners' decisions.

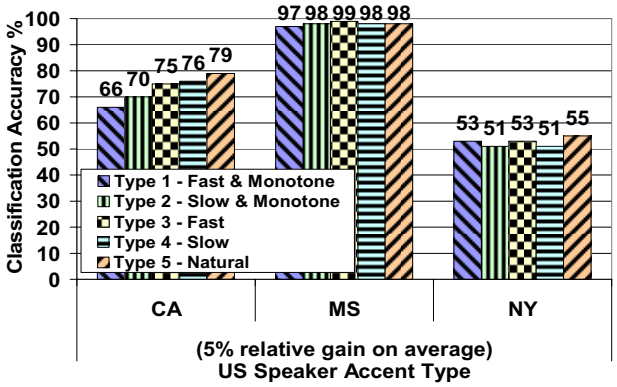

Figure 6. The Effect of the Presence of the original Tempo and/or f0 Contour on Perceptual Accent with Intelligible (Type 2) Speech. The effect is most clearly observed in the case of CA accent – the accuracy for Type 1 (processed fast tempo and flat f0 contour) speech is 66%, and Type 5 (unprocessed, original) speech is 79%.

6. Discussion

In the area of automatic accent classification for automatic speaker identification and ASR engines, it is important to establish human performance in order to assess the significance

of automatic systems, and to provide scientific insights for further advancements in speech technology. The two main observations[‡] from our results are as follows: ***(1)*** prosody (f0 contour and tempo in this study) contributes to perceptual accent classification, and ***(2)*** accent perception involves both bottom-up (e.g., acoustic input) and top-down (e.g., mental representation of prototypical accent characteristics) processing, which can be multi-dimensional. That is, the cognitive process of accent classification is based not only on pronunciation characteristics but also on other factors such as utterance content via transcripts or speech signal. This is the case even though this experiment focuses on accent/pronunciation variation and excludes other dialectal variability such as word selection. This may mean that accent classification by human perception requires somewhat meaningful, comprehensible speech samples (e.g., word, phrase, sentence, etc.), which are not necessarily the case for automatic systems.

Multidimensionality of accent perception also brings up important issues such as the relationship between listener-accent background and speaker-accent type (e.g., a listener's familiarity with the speaker's accent), which would affect the authenticity of the conceptual representation of prototypical accent characteristics, and therefore the accuracy of perceptual classification. As Fig. 7 illustrates, when unprocessed, natural speech was presented, listeners from all regions in the US were able to classify Mississippi accent correctly (98% to 100%). However, listeners from the southern US were not able to classify California accent and New York accent as well as listeners from western and northern US (50% vs. 73% and 79% in the case of California accent). Since the number of listeners from each region in this study was not equally distributed, this number may not provide strong indications. However, this result still points to one dimension involved in accent perception and suggests its importance.

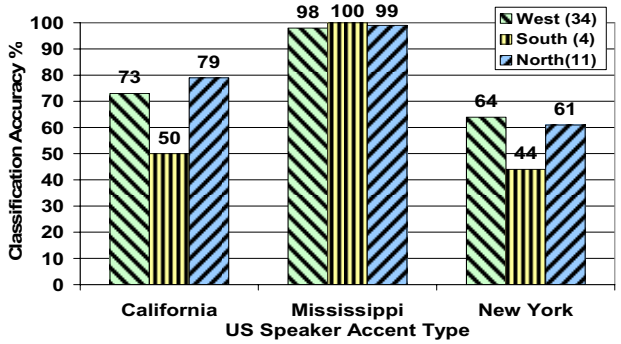

Figure 7. The Effect of Listener Accent Background. For example, the listeners from western US accurately classified California accent 73%, Mississippi accent 98%, and New York accent 64% of the time.

Another dimension which needs to be considered in accent perception is content of the speech, as the results in this study indicated. Although prosody (alone or together with vowel and accent characteristics) can provide meaningful cues for perceptual accent classification, without understanding what has been said either through speech signal or through transcripts, listeners were not able to utilize the information effectively. In our previous study[13], the results indicated that comprehensibility of the speech (measured by transcription accuracy in this case) influenced the classification accuracy and confusability among different accents, especially for listeners who were unfamiliar with accents under test. Accent with higher comprehensibility was less often confused with accent that is less comprehensible. Accents that were similarly comprehensible were more often confused with each other.

Together with the observations from our previous study[13], the results in this present study point to the complexity of the cognitive process involved in accent classification. It involves bottom-up and top-down processing, and those processes are multi-dimensional and not limited to speech production characteristics such as vowel and consonant characteristics and prosody. This suggests the importance of further understanding of the cognitive process involved in accent classification in order to incorporate human and machine performance to achieve highest overall accuracy in accent classification and identification.

7. Conclusion

Our goal in this study has been to provide perceptual assessment of accent variations for accent identification applications. This is important in order to establish benchmark human performance to compare with automatic systems as well as to leverage perception and algorithms most effectively to achieve highest performance for both security and communication applications. Main trends indicated that (1) f0 contour and tempo characteristics contribute to accent perception, (2) listeners' perception is influenced by the availability of utterance content (e.g., meaning, word and phoneme sequence) via intelligible speech or via transcripts when the speech is unintelligible, and (3) listeners' decisions are affected by conceptual representation of prototypical accent characteristics. These observations suggest that listeners use both bottom-up (e.g., speech signal input) and top-down (e.g. conceptual representation of prototypical accent characteristics) processing, and it is multi-dimensional. That is, not only pronunciation characteristics but also availability of utterance content influences accent perception, although this experiment focused only on pronunciation characteristics and did not include word use that are dialect dependent. Overall, our analyses point to the importance of understanding cognitive aspects of accent variation, which will contribute to further development of speech technology, including automatic accent and speaker identification, and speech recognition.

[‡] Due to limited space, statistical analysis results are not presented in this paper. More details are described in Ikeno[12].

REFERENCES

[1] Arslan, L., Hansen, J.H.L (1996). Language Accent Classification in American English, *Speech Communications*,18(4), pp.353-367.

[2] Yanguas, L.R., O'Leary, G.C., Zissman, M.A., Incorporating Linguistic Knowledge into Automatic Dialect Identification of Spanish, *ICSLP-98.*

[3] Angkititrakul P. and Hansen J.H.L. (2006). Advances in Phone-Based Modeling for Automatic Accent Classification. *IEEE Trans. Speech & Audio Proc., vol. 14(2):634-646, March 2006.*

[4] Crystal D. (2003). A dictionary of linguistics and phonetics. Malden, MA: Blackwell Pub.

[5] Clopper C.G.(2004). *Linguistic Experience and the Perceptual Classification of dialect Variation*. Ph.D. Dissertation, Indiana University.

[6] Clopper C.G., and Pisoni D.B. (2004b). Some Acoustic Cues for the Perceptual Categorization of American English Regional Dialects. *Journal of Phonetics.* vol.32, pp.111-140.

[7] Flege, J.E. (1988). Factors affecting degree of perceived foreign accent in English sentences. J. Acoust. Soc. Amer., v.84, p.70-77.

[8] Jilka M. (2000).The contribution of intonation to the perception of foreign accent. Stuttgart:PhD thesis, University of Stuttgart.

[9] Munro M.J., Derwing T.M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning vol.49, pp.285-310(26).

[10] Moulines, E. and Laroche, J. (1995). Non-parametric techniques for Pitch-Scale and Time-Scale Modification of Speech. *Speech Communication*, vol.16, pp.175-205.

[11] Moulines, E. and Sagisaka, Y. (1995). Voice Conversion: State of the Art and Perspectives. *Speech Communication*, vol.16(2).

[12] Ikeno, A. (2005). *Perceptual Cues in English Accent Variation: The Role of Prosody and Listener Accent Background*, Doctoral Dissertation, University of Colorado at Boulder.

[13] Ikeno, A., Hansen, J.H.L. (2006). "Perceptual Recognition Cues in Native English Accent Variation: "Listener Accent, Perceived Accent, and Comprehension," *IEEE ICASSP-06*, vol. 1, p401-404, Toulouse, France.