Automatic Detection of “g-dropping” in American English Using Forced Alignment

Jiahong Yuan [1], Mark Liberman [2]

University of Pennsylvania

[1] jiahong@ling.upenn.edu

[2] myl@cis.upenn.edu

***Abstract*—This study investigated the use of forced alignment for automatic detection of “g-dropping” in American English (e.g., walkin'). Two acoustic models were trained, one for *-in'* and the other for *-ing*. The models were added to the Penn Phonetics Lab Forced Aligner, and forced alignment will choose the more probable pronunciation from the two alternatives. The agreement rates between the forced alignment method and native English speakers ranged from 79% to 90%, which were comparable to the agreement rates among the native speakers (79% - 96%). The two variations of pronunciation not only differed in their nasal codas, but also – and even more so – in their vowel quality. This is shown by both the KL-divergence between the two models, and that native Mandarin speakers performed poorly on classification of “g-dropping”.**

I. Introduction

The term “g-dropping” refers to the phenomenon in English where the *-ing* ending is pronounced with an alveolar nasal [n]. It is shown in the conventional orthography by the use of an apostrophe in place of *g*, as in *walkin’* and *nothin’*.

“g-dropping” has been extensively studied in the field of sociolinguistics since the 1950s, commonly termed as the ING variable (e.g. [1]-[7]). These studies have found that “g-dropping” is associated with both lower socioeconomic status and informal speech style. In North America the Southern speakers have a higher rate of “g-dropping” than the Northern speakers, and to a less certain degree, the younger and male speakers tend to have a higher rate of “g-dropping” than the older and female speakers, respectively.

The phonological and syntactic constraints on “g-dropping” have also been investigated. “g-dropping” only occurs in unstressed environments but not in stressed syllables (such as in the word *sing*). The words *everything* and *anything* have much lower “g-dropping” rates than *something* and *nothing*, which was accounted for by stress assignment in [4], i.e., *everything* and *anything* have a secondary stress on *-ing* but *something* and *nothing* have not. Reference [8] reported the syntactic constraints on the rate of “g-dropping” as follows: most in progressives and participles, less in adjectives, even less in gerunds and least of all in nouns like *ceiling* and *morning*.

The phonetic realizations of the ING variable are rarely studied [9]. Generally speaking, the *-in’* form may be realized as [ɪn], [ən], or syllabic [n] whereas *-ing* is [ɪŋ]. There has been an extensive discussion on Language Log about the vowel quality difference between the two forms (e.g. [9], [10]).

Previous studies on “g-dropping” have relied on impressionistic coding of the variable. Although it is a relatively easy task for native speakers to manually classify “g-dropping” [11], it is expensive, inconsistent and impractical when we move on from analysing small datasets to thousands of hours of speech that are now openly available. In this study we investigate the use of forced alignment for automatic classification of ‘g-dropping’. We evaluate the method by comparing the results with native speakers’ manual coding. Finally, we explore the acoustic difference between the two variations (*-in’* and *-ing*) through both computing the divergence between the two models, and comparing native English and Mandarin speakers’ performance on their classification of “g-dropping”.

II. Data, Method and Evaluation

We trained two acoustic models, one for *-in'* (/IHN/) and the other for *-ing* (/IHNG/). The models were GMM-based, five-state HMMs on 39 PLP coefficients [12]. The parameters of the models were initially estimated using the Buckeye Corpus, which contains the speech of 40 speakers conversing freely with an interviewer [13]. The corpus provides detailed phonetic transcription. Based on the phonetic transcription, 5,816 *-ing/-in’* words in the corpus were selected and used in our study, excluding the *-ing*s in stressed syllables or in other phonetic forms such as in *gonna*. 23% of the 5,816 words are “dropping g’s” (*-in'*). The “g-dropping” rates range from 0.02 to 0.61 among the 40 speakers; and the distribution is shown in Figure 1. Table I lists the “g-dropping” rates of the most frequent words in the corpus.

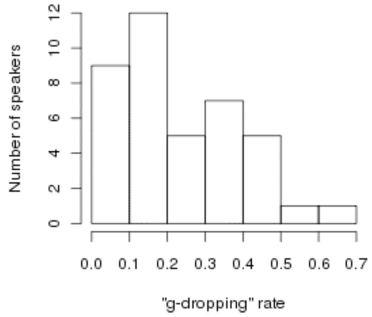

Fig. 1. Histogram of the “g-dropping” rates of the 40 speakers in Buckeye.

TABLE I
"G-DROPPING" RATES OF MOST FREQUENT WORDS IN BUCKEYE

Word	Frequency	"g-dropping" rate
something	454	0.278
going (excluding *gonna*)	367	0.491
doing	305	0.361
everything	305	0.010
anything	293	0.003
being	220	0.282
getting	143	0.084
having	120	0.233
working	120	0.308
saying	115	0.417
trying	111	0.712
nothing	96	0.260
talking	85	0.365
looking	83	0.277
coming	80	0.388
interesting	80	0.038

We randomly selected 200 words, 100 with "g-dropping" and 100 without "g-dropping", to form a test set. The rest were used for training. The training was done using the HTK toolkit [14].

The models trained on the Buckeye Corpus were then re-estimated using the much larger SCOTUS corpus and the CMU pronouncing dictionary ([15], [16]). The SCOTUS corpus includes more than 50 years of oral arguments from the Supreme Court of the United States. 34,656 speaker turns from the arguments of the 2001 term were used for the re-estimation, which is the same dataset we have used to train the Penn Phonetics Lab Forced Aligner [17]. The SCOTUS corpus only has word transcriptions but no phonetic transcriptions. We edited the CMU dictionary to include two variations of pronunciation for each *-ing* word, one with /IHN/ and the other with /IHNG/. /IHN/ and /IHNG/ were treated as unitary phonemes. During each iteration of training, the 'real' pronunciations of the *-ing*s were automatically determined, and then the acoustic models of /IHN/ and /IHNG/ were updated.

As a preliminary analysis of the Justices' speech, Table II lists the "g-dropping" rates of the speakers in the SCOTUS corpus, determined by the final models trained on the corpus.

TABLE II
"G-DROPPING" RATES OF SPEAKERS IN SCOTUS

Speaker	Number of *-ings*	"g-dropping" rates
Breyer	884	0.196
O'Connor	292	0.086
Ginsburg	781	0.168
Kennedy	372	0.543
Rehnquist	303	0.389
Scalia	1,096	0.193
Souter	836	0.194
Stevens	298	0.279
Others (lawyers, etc.)	6,843	0.241

The two models were added to the Penn Phonetics Lab Forced Aligner, and in the test stage forced alignment will choose the more probable pronunciation from the two alternatives for the acoustic observation. The procedure is shown in Figure 2.

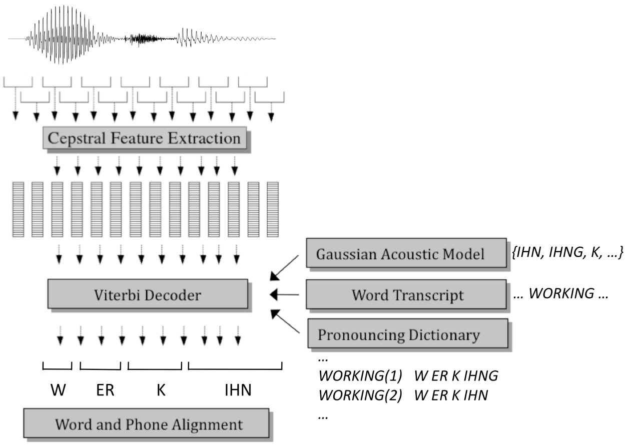

Fig. 2. Procedure for "g-dropping" detection using forced alignment.

We tested the method on the 200 words randomly selected from the Buckeye Corpus. In addition, we did a forced-choice perception experiment using the same 200 words. In the experiment, the sentences containing the target words were presented to the subjects using Praat, along with the word transcription except the target word, as shown in Figure 3. The subjects were asked to judge whether the untranscribed target word is "g-dropped" or not. They could listen to the sentences and words as many times as they like. Eight native speakers of American English participated in the experiment.

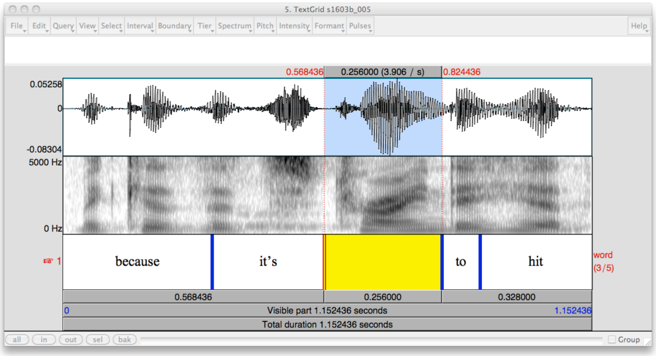

Fig. 3. The interface presented to the subjects in the perception experiment.

The pairwise percentage agreements among the forced aligner and the humans, including both the eight subjects and the Buckeye transcription, are listed in Table III. We can see from the table that the agreement rates between the forced alignment method and the humans ranged from 79% to 90% (mean = 0.849). They are comparable to the agreement rates among the humans, which ranged from 79% to 96% (mean = 0.863).

TABLE III
PAIRWISE PERCENTAGE AGREEMENTS AMONG HUMANS AND FORCED ALIGNMENT (BU=BUCKEYE, AL=FORCED ALIGNMENT, S=SUBJECT)

	Bu	S1	S2	S3	S4	S5	S6	S7	S8	Al
Bu	1.0	.87	.84	.81	.84	.79	.90	.86	.84	.88
S1	.87	1.0	.92	.84	.96	.85	.96	.93	.89	.88
S2	.84	.92	1.0	.84	.93	.85	.93	.92	.87	.86
S3	.81	.84	.84	1.0	.86	.82	.86	.86	.79	.79
S4	.84	.96	.93	.86	1.0	.88	.94	.94	.87	.85
S5	.79	.85	.85	.82	.88	1.0	.87	.88	.83	.80
S6	.90	.96	.93	.86	.94	.87	1.0	.93	.89	.90
S7	.86	.93	.92	.86	.94	.88	.93	1.0	.89	.85
S8	.84	.89	.87	.79	.87	.83	.89	.89	1.0	.83
Al	.88	.88	.86	.79	.85	.80	.90	.85	.83	1.0

III. ACOUSTIC CHARACTERISTICS OF "G-DROPPING"

In this section we study the acoustic difference between the two variations of pronunciation of the ING variable. We first compare native English and Mandarin speakers' performance on classifying "g-dropping", and then we compute the KL-divergence of the two acoustic models for *in'* and *ing*.

Mandarin Chinese has both alveolar and velar nasal codas, which is similar to English. On the other hand, there is no lax vowel /ɪ/ in Mandarin Chinese. If the two variations of pronunciation of the ING variable in English are different from each other mainly on their codas, we expect that native speakers of Mandarin Chinese will be able to perform reasonably well on coding the variable. If, however, the difference is mainly realized on the vowel, we expect that Chinese speakers will have difficulty in classifying "g-dropping" in English, because of the lax vowels in *-in'* and *-ing*. Ten native speakers of Mandarin Chinese participated in the experiment. They were graduate students at University of Pennsylvania, and spoke English as their second language fluently or near-fluently.

To compare the performance of Mandarin Chinese and American English listeners, we used the majority vote of the English listeners' results as the gold standard. Figure 4 draws the classification accuracies of the Chinese and English listeners, as well as that of the forced alignment method. We can see that the accuracies of the Chinese listeners were significantly lower than both the English listeners and the alignment method. Figure 5 draws the true and false positive rates of the listeners, in which the presence of "g-dropping" was treated as the positive. The figure also shows that Mandarin Chinese listeners performed poorly on classifying "g-dropping" in English.

The two acoustic models for *-ing* and *-in'* are GMM-based, five-state HMMs. We computed the distance of the GMM models at each of the HMM states. A natural measure between two distributions is the Kullback-Leibler divergence [18]; however, it cannot be analytically computed in the case of a GMM. We adopted a dissimilarity measure proposed in [19], which is an accurate and efficiently computed approximation of the KL-divergence. Figure 6 shows the distance between the two models at each HMM state. We can see that the distance reaches its peak at the middle state, and it is larger on the left side (the vowel side) than the right (the nasal coda side). This result suggests that the two variations of pronunciation of the ING variable are more different in their vowels than in their nasal codas.

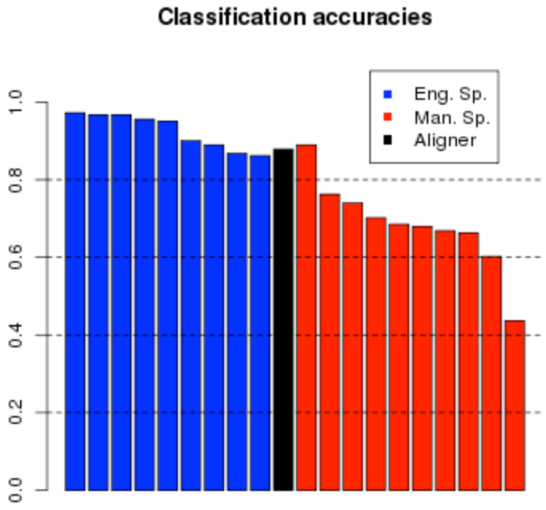

Fig. 4. Classification accuracies of English listeners, Chinese listeners, and the alignment method, using the majority vote of the English listeners as the gold standard.

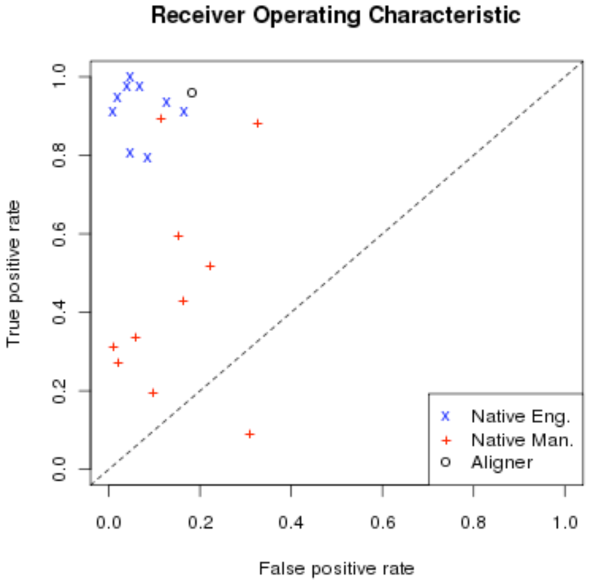

Fig. 5. True and false positive rates of English listeners, Chinese Listeners, and the alignment method, using the majority vote of the English listeners as the gold standard and the presence of "g-dropping" as the positive.

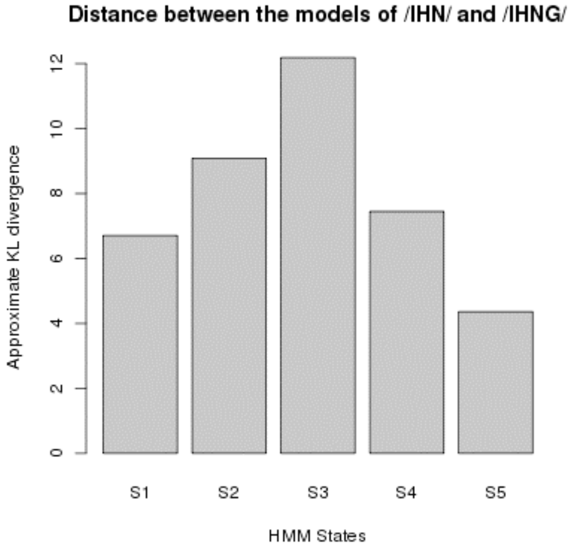

Fig. 6. KL-divergence between the acoustic models for *-ing* (/IHNG/) and *-in'* (/IHN/) at each of the five HMM states.

IV. Conclusions

This study investigated the use of forced alignment for automatic detection of "g-dropping" in American English. Two acoustic models were trained, one for *-in'* and the other for *-ing*. The acoustic models were GMM-based, five-state HMMs. The parameters of the models were initially estimated using the Buckeye Corpus and then re-estimated using the SCOTUS corpus. The models were added to the Penn Phonetics Lab Forced Aligner, and forced alignment will choose the more probable pronunciation from the two alternatives for the acoustic observation.

We tested the method on 200 word tokens randomly selected from the Buckeye Corpus. In addition, we did a forced-choice perception experiment using the same 200 words. Eight native American English speakers and 10 native Mandarin Chinese speakers participated in the experiment. The agreement rates between the forced alignment method and the native English speakers ranged from 79% to 90% (mean = 0.849), which were comparable to the agreement rates among the native English speakers (79% to 96%, mean = 0.863). Our study also demonstrated that the two variations of pronunciation not only differed in their nasal codas, but also – and even more so – in their vowel quality. This is shown by both the KL-divergence between the two models, and that native Mandarin speakers performed poorly on classification of "g-dropping" in English.

We are currently testing the robustness of the method on a variety of large speech corpora, including political speech. Our preliminary results on President Obama's weekly addresses have been very encouraging, at about 90 percent agreement with the second author of the paper.

Acknowledgment

We thank Jonathan Izak and Nicole Wang for their help running the perception experiment. This work is supported in part by NSF grant IIS-0964556.

References

[1] J. Fischer, "Social influence of a linguistic variant," *Word*, 14, 47–56, 1958.
[2] W. Labov, *The Social Stratification of English in New York City*, Washington, DC: Center For Applied Linguistics, 1966.
[3] P. Trudgill, 1974. *The Social Differentiation of English in Norwich*, Cambridge: Cambridge University Press, 1974.
[4] A. Houston, *Continuity and Change in English Morphology: The Variable (ING)*, PhD dissertation, University of Pennsylvania, 1985.
[5] W. Labov, *Principles of linguistic change: Social factors*, New York: Blackwell, 2001.
[6] K. Campbell-Kibler, *Listener Perceptions of Sociolinguistic Variables: The Case of (ING)*, PhD dissertation, Stanford University, 2006.
[7] K. Hazen, "A vernacular baseline for English in Appalachia," *American Speech*, 83, 116-140, 2008.
[8] W. Labov, "The child as linguistic historian," *Language Variation and Change*, 1, 85–98, 1989.
[9] M. Liberman, "Symbols and signals in g-dropping," Language Log, Available: http://languagelog.ldc.upenn.edu/nll/?p=3037, 2011.
[10] M. Liberman, "Pawlenty's linguistic "southern strategy"?" Language Log, Available: http://languagelog.ldc.upenn.edu/nll/?p=3032, 2011.
[11] K. Hazen, "The in/ing variable," In K. Brown, (ed.), *Encyclopedia of Language and Linguistics*, pp. 581-584, Cambridge: University of Cambridge, 2005.
[12] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, 1990.
[13] M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech (2nd release)* [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor), 2007.
[14] http://htk.eng.cam.ac.uk/
[15] http://www.oyez.org/
[16] http://www.speech.cs.cmu.edu/cgi-bin/cmudict/
[17] http://www.ling.upenn.edu/phonetics/p2fa/
[18] S. Kullback, *Information theory and statistics*, New York: Dover Publications, 1968.
[19] J. Goldberger, and H. Aronowitz, "A distance measure between GMMs Based on the unscented transform and its application to speaker recognition," *Proceedings of Interspeech '05*, pp. 1985-1989, 2005.