IMPLICATIONS OF GLOTTAL SOURCE FOR SPEAKER AND DIALECT IDENTIFICATION*

Lisa R. Yanguas and Thomas F. Quatieri
M.I.T. Lincoln Laboratory
Lexington, MA 02420-9185
lryangu | tfq @sst.ll.mit.edu

Fred Goodman
Autometric, Inc.
Springfield, VA 22153

ABSTRACT

In this paper we explore the importance of speaker specific information carried in the glottal source. We time align utterances of two speakers speaking the same sentence from the TIMIT database of American English. We then extract the glottal flow derivative from each speaker and interchange them. Through time alignment and this glottal flow transformation, we can make a speaker of a northern dialect sound more like his southern counterpart. We also time align the utterances of two speakers of Spanish dialects speaking the same sentence and then perform the glottal waveform transformation. Through these processes a Peruvian speaker is made to sound more Cuban-like. From these experiments we conclude that significant speaker and dialect specific information, such as noise, breathiness or aspiration, and vocalization, is carried in the glottal signal.

1. INTRODUCTION AND MOTIVATION

In the area of speaker identification, researchers have questioned the extent to which speaker specific information is carried in the glottal flow [1] versus the vocal tract. Earlier work presented an automatic technique for estimating and modeling the glottal flow derivative waveform from speech and applied the model parameters to a speaker identification task [2]. For a large TIMIT data subset, averaging over both male and female speaker identification scores, a combination of both coarse- and fine-structure glottal features was shown to contain significant speaker-dependent information and yielded an approximate 30% error rate in the identification task. When used in combination with a traditional speaker identification system based on mel-cepstral measures on a subset of the NTIMIT database, the glottal flow estimation and modeling technique reduced the error rate by about 5%. In this paper we approach the glottal flow (source) vs. vocal tract (system) issue using speech analysis/synthesis as a means toward gaining insight into which measurements are most informative in the speaker identification task. However, we also show important ramifications for and insights into dialect identification. The work suggests further study in linking speaker and dialect identification more closely.

2. TIME ALIGNMENT

We began by looking at utterances spoken by two speakers from the TIMIT database of dialects of American English. Both speakers uttered the same shibboleth sentence ("She had your dark suit in greasy wash water all year.") One was a speaker of a northern (i.e. New York City) dialect, while the other was a southern dialect speaker. Both speakers were males. We first time aligned the utterances of the two speakers using a time-scaling algorithm based on sinusoidal analysis/synthesis [3]. A nonuniform time scale mapping was derived by first computing the ratio of time durations of matched phonemes, (phoneme timings given within the TIMIT database). Using a 10 ms frame, this phoneme-based time mapping was then converted to a frame-based mapping for use in the sinusoidal analysis/synthesis system. As might be expected, the southern speaker's rate of speaking was significantly slower than the northerner's, so this step alone had the effect of elongating the northerner's speech to align in time with that of the southerner. An expert linguist, upon comparing the speech files following the time scaling with the original speech, assessed the northerner as sounding more "southern-like," strictly from the time alignment procedure. Thus, the result of the time alignmemt step has implications for the use of supra-segmental methods in dialect identification.

3. GLOTTAL FLOW DERIVATIVE

The glottal airflow volume velocity, denoted by u(t), acts as the source, sometimes also referred to as the "excitation," to the vocal tract. The vocal tract impulse response is denoted by h(t). The volume velocity output of the vocal tract is then modified by the lip impedance. Because the pressure/volume velocity relation at the lips can be approximated by a differentiator, the speech pressure waveform s(t) measured in front of the lips can be expressed approximately as

$$s(t) = d[u(t)*h(t)]/dt = [du(t)/dt]*h(t) \qquad (1)$$

where "*" denotes convolution. The effect of radiation is typically included in the source function; as in (1), the source to the vocal tract, therefore, becomes the derivative of the glottal flow volume velocity, henceforth denoted by v(t)= du(t)/dt. In obtaining the glottal flow derivative, applying the lip radiation effect to the source flow, rapid closing of the vocal folds results in a large negative impulse-like response at glottal closure, called the glottal pulse. In voiced speech, the time interval during which the vocal folds are closed, and during which no flow occurs, is referred to as the glottal closed phase. The time interval over which there is nonzero flow and the vocal folds are fully or partially open is referred to as the glottal open phase, while the time interval from the most negative value of the glottal flow derivative to the time of glottal closure is referred to as the return phase.

The theory for the production of voiced speech suggests that an accurate estimate of the vocal tract response h(t) can be calculated

THIS WORK WAS SUPPORTED BY THE DEPARTMENT OF THE AIR FORCE. OPINIONS, INTERPRETATIONS, CONCLUSIONS AND RECOMMENDATIONS ARE THOSE OF THE AUTHORS AND NOT NECESSARILY ENDORSED BY THE UNITED STATES AIR FORCE.

during the glottal closed phase when there is no source -vocal tract interaction [4,5]. This estimate can then be used to inverse filter the speech signal. For quasi-periodic speech signals, inverse filtering results in an estimate of the glottal flow derivative v(t) consisting of a coarse-structure component giving the general flow and a fine-structure component including aspiration and a perturbation in the flow referred to as "ripple." Aspiration is sometimes manifested or perceived as "noisiness" or "breathiness" and may occur anywhere in the glottal cycle. Ripple is associated with first-formant modulation and is due to the time-varying and nonlinear coupling of the source and vocal tract cavity typically during the open phase [4,5]. In unvoiced speech, inverse filtering results in noise or an impulse-like source during fricative or plosive sounds, respectively.

The glottal flow derivative estimate is obtained by inverse filtering the speech waveform with a vocal tract filter derived over a glottal closed-phase estimate, according to a "stationary" region of formant modulation [2]. In finding a closed-phase estimate, we first find an approximate location of the glottal pulse by peak-picking an excitation estimate obtained by an initial pass at inverse filtering the speech waveform. This glottal pulse location is used as an anchor point to define a region over which formant modulation is computed via the sliding covariance method of linear prediction. Statistics are derived, over this region, on the formant modulation function for determining a closed-phase estimate. Formant modulation can be exploited for this purpose because a formant change occurs in going from the glottal closed phase to open phase, in which source/vocal tract interaction occurs. Because of the presence of aspiration and because of the nonlinear nature of the source-filter interaction, however, the formants will vary somewhat even with a constant glottal area, as for example during the closed phase of a breathy speaker. When the glottis begins to open, the formants will move from the relatively stable values they maintained during the closed phase. An interval of relatively small variance defines the glottal closed phase over which the vocal tract filter is estimated by the covariance method of linear prediction. The resulting vocal tract estimate is then used to perform a second pass of inverse filtering on the original waveform. This analysis method provides an accurate vocal tract filter estimate, as well as glottal flow derivative estimate on each pitch period during voicing. An example of glottal flow derivative estimates obtained from the same phone of a northern and southern speaker (from the same time-aligned TIMIT text) are shown in Figure 1. Aspiration and ripple contributions and a typically large glottal pulse are seen in the flow derivative estimates.

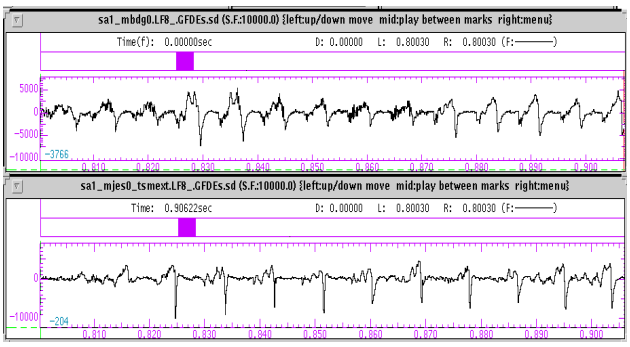

Figure 1: Glottal flow derivative displays from the utterance "She had your dark suit in greasy wash water all year." for time-aligned southern (top) and northern (bottom) dialect speakers.

4. GLOTTAL FLOW INTERCHANGE

Earlier experimentation [1, 2] has shown that speaker-specific information is carried in the glottal source. Thus, removing the glottal component from the utterance of a speaker A, and replacing it with that of another speaker B, should result in a signal sounding more like speaker B. In performing this experiment, we show how this is borne out with respect to two speaker pairs each of English and Spanish, but how it also hinges on the salient characteristics of the particular speakers and/or dialects under consideration. For example, previous work in dialect identification in Spanish [6] has shown that aspiration, for example, is an important component with respect to discrimination. We will show how this distinction can be exploited for our work here with glottal flow transformations.

4.1 Interchange method

Our glottal flow derivative estimation method provides both the flow derivative and vocal tract filter estimates on each pitch period. One approach to interchange the glottal flow derivatives of two speakers under consideration is to simply interchange the flow function on each pitch period. However, because the pitch contours of the two speakers differ, a common frame normalization is invoked whereby one vocal tract filter and one glottal flow derivative estimate are convolved on a fixed (10 ms) frame and combined over successive frames with overlap-add synthesis. On each frame, speaker A retains his/her vocal tract response, but uses speaker B's pitch and glottal flow excitation.

4.2 Interchanging the glottal flow for American English dialect speakers

We computed the glottal flow derivative waveform for the northern and the southern dialect TIMIT speakers for the utterance "She had your dark suit in greasy wash water all year.", a segment of which was given in Figure 1. By studying the glottal waveforms and the spectrograms of the respective glottal sources we observed that the southern glottal source appeared significantly noisier. Figure 1 illustrates the relatively larger aspiration component within the Southern glottal waveform. This was confirmed upon listening to both the glottal and speech waveforms, as the southern waveforms sounded significantly more aspirated than that of the northerner.

We then added the southern glottal waveform to the vocal tract information for the northern speaker. By default, pitch was also interchanged during the process. For clarity here we will identify the dialect with the vocal tract information; thus, the northern speaker refers to the vocal tract information for the northerner, who then receives the glottal waveform from the southern speaker. Following the transformation, the northern speaker has acquired more "noise" or aspiration and is much closer to the original southern speaker. Figure 2 illustrates the effect of replacing the glottal flow derivative of the northern speaker with that of the southern speaker. The exchange of pitch information likely contributes to this increased similarity, as well. In fact, in a separate experiment the pitch of the northern speaker was replaced with his southern counterpart, resulting in a move toward the southern characteristics, but this was not as significant a change as when the entire glottal derivative waveform was replaced. The pitch transformation was performed using the sinusoidal analysis/synthesis system [3]. A pitch mapping was

derived by first computing the pitch of both speakers and then finding the ratio of their pitch contour at a 10 ms frame interval. The pitch mapping was invoked on the time-aligned northern speaker during time regions in which both speakers were declared voiced.

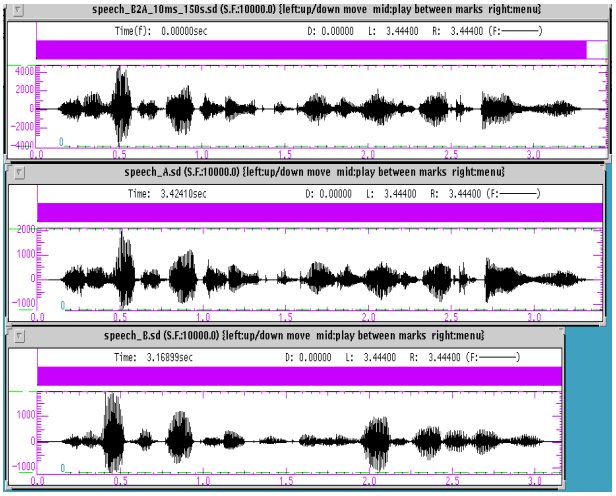

Figure 2: Waveforms displays of the utterance “She had your dark suit in greasy wash water all year.” for modified northern (top), having the glottal flow and time-alignment of the original southern (center). The original northern (bottom) is also shown.

5. LINGUISTIC ANALYSIS

There is still some distortion in the recreated speech files that tends to interfere with human perception, due to unvoiced/voiced timing mismatches and occasional inaccurate vocal tract response estimates. Nevertheless, an expert linguist characterized the northern dialect speaker as sounding “more southern-like” following this glottal transformation. This was somewhat less compelling on an overall phrase level due to the interference from this type of distortion; in contrast, however, comparisons at both the individual word and the phone level showed that the similarities were rather dramatic and quite convincing.

Speakers of southern dialects of American English have been noted to have very marked pronunciations of vocalic segments, such as vowels and diphthongs. Vocalic segments are voiced by definition in English, and thus involve vibration of the vocal chords and the engagement of the glottis. In comparing the spectrograms depicted in Figure 3 below, we observe that the northerner’s spectrogram looks more clearly defined than the southerner’s, which shows evidence of more noise in the southern speaker’s utterance as well as a longer duration for the vocalic segment. Some of this information is carried in the glottal waveform, as well, causing us to surmise that when the glottal waveform transformation is performed, the northern speaker sounds decidedly more southern-like with respect to particular words within which this phenomenon is at work. These will include words with dialectically characteristic vowel segments such as “wash,” “dark,” and “water.”

The southern dialect speaker also voices the [s] in “greasy,” thereby pronouncing it like a [z]. The voicing of unvoiced intervocalic consonants is an observed characteristic of southern dialect speech. Here it serves to lend additional credence to our notion of aspiration or noisiness being conveyed through the glottal waveform. While [s] and [z] are both fricatives and thus strident sounds, the voiced quality or increased glottal vibration of the [z] adds to the perception of noisiness or breathiness (i.e. aspiration) with respect to the Southern speaker (see Figure 4).

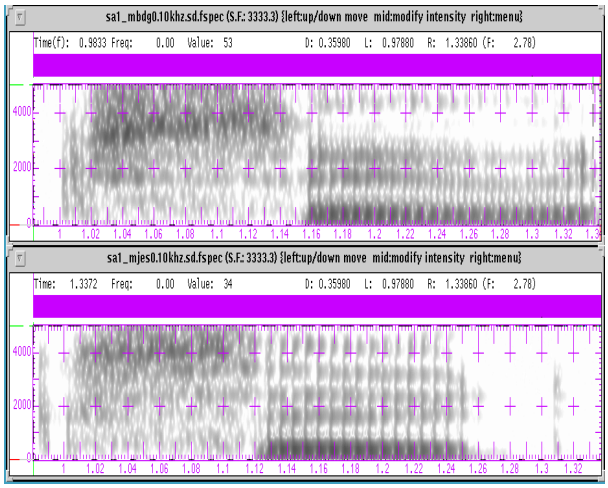

Figure 3: Wideband spectrograms of the word “suit.”: northern speaker (bottom) vs. southern speaker (top).

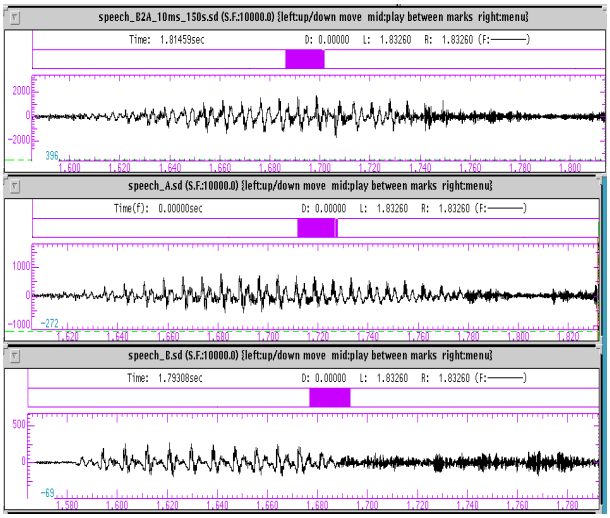

Figure 4: Waveform displays of “greasy” from “She had your dark suit in greasy wash water all year.” for modified northern (top), having the glottal flow and time-alignment of the original southern (center). The original northern (bottom) is also shown.

6. IMPLICATIONS FOR DIALECT IDENTIFICATION

The foregoing analysis suggests obvious implications for dialect identification. Admittedly, the data set is extremely small; thus far and we have focused only on one speaker from each of two dialects of American English. Nevertheless we were sufficiently encouraged to expand our study and explore whether similar precepts would apply to dialects of another language.

We attempted this same procedure on two speakers of two different dialects of Spanish. We chose two male speakers from the Miami Corpus of Spanish Dialects Database [7]. One was a speaker of Cuban Spanish and one of Peruvian. They were both speaking the same shibboleth utterance ("El ajo es muy fuerte" - ''Garlic is very strong."). We again began by time aligning the two utterances. In this case the Peruvian speaker's rate of speech was discernibly faster than that of the Cuban. In keeping with our discoveries with respect to English, the time alignment procedure, alone, served to make the Peruvian sound more Cuban-like, based on an expert linguist's assessment.

In previous work analyzing the Miami Corpus data [6,7,8], it was shown that Cuban Spanish is more heavily aspirated than Peruvian Spanish. Cuban speakers are perceived as "breathier" and tend to aspirate various phones, notably [s] in word- or morpheme-final position during rapid speech and during voicing. It was this distinction that we hoped to exploit in performing the glottal waveform transformation. Figure 5 compares glottal flow derivative estimates from a Peruvian and Cuban speech segment. As anticipated, the Cuban display shows more noise/aspiration.

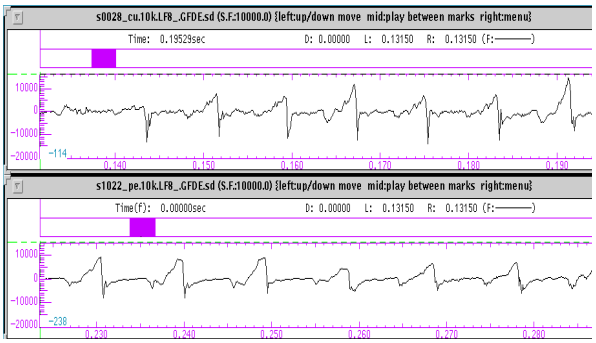

Figure 5: Glottal flow derivative displays from the utterance "El ajo es muy fuerte" - ''Garlic is very strong." - for Cuban (top) and Peruvian (bottom) dialect speakers.

When the Peruvian speech is time aligned to match that of the slower rate of the Cuban speech, it begins to sound more like the slower paced utterance. The native speaker/linguist's assessment was also that after the additional glottal transformation, the Peruvian speaker sounded even more Cuban-like. As in the previous experiment, a pitch transformation alone did not result in as effective a change as did the complete glottal flow replacement. This parallels our earlier results on English between northern and southern dialect speakers. The glottal waveform carries a good deal of information, particularly relating to speech that characteristically has pronounced aspiration, noise or breathiness plus vocalization. These types of glottal waveform transformations produce the effect of causing the "recipient" speech to sound more like the "donor" speech.

7. CONCLUSIONS AND FUTURE WORK

Our initial intent was to look at glottal flow transformations in order to gain a better understanding of and improvement in the speaker identification task. While these experiments are rather preliminary on only two pairs of speakers, it seems clear that in addition to ramifications for speaker identification, glottal waveform transformations have obvious implications for dialect identification. In our first-pass experiments with both English and Spanish dialects, results map well with respect to specific information carried in the glottal waveform and exchanged during the transformation process. Duration and pitch are clear contributors to speech perception, but elements such as aspiration, noise, or breathiness, stridency, and voicing or vocalization all play a role in perception and analysis of voice, as well. Based on these findings, it appears that although dialect identification has usually been approached in a manner similar to language identification, it may actually be more closely related to the speaker identification task. We plan to explore other speaker identification approaches to dialect identification in the hopes of improvement in performance. We have established a firm baseline with respect to two pairs of speakers of dialects of American English and Spanish in terms of glottal transformation. We foresee expanding our data set and testing our procedures on a variety of speakers, as well as on a larger dialect set. In addition, refinements to the algorithm, which would diminish distortion currently impeding perception, will help greatly in our voice conversion processes. Finally, we plan to delve further into the transformation process and ideally merge positive results on this task with results from more traditionally-based speaker and dialect identification systems to thereby improve overall performance and accuracy rates.

Acknowledgment: The authors thank Marc Zissman for detailed comments on the manuscript.

REFERENCES

[1] D.G. Childers. Glottal Source Modeling for Voice Conversion. in Speech Communication, volume 16, pages 127-138, 1995.

[2] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds. Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification. Submitted to IEEE Transactions on Speech and Audio.

[3] T.F. Quatieri and R.J. McAulay, Shape-Invariant Time-Scale and Pitch Modification of Speech, IEEE Trans. on Signal Processing, Vol. 40, No. 3, March 1992.

[4] G. Fant. Glottal flow: models and interaction. *Journal of Phonetics*, 14:393-399, 1986.

[5] T.V. Ananthapadmanabha and G. Fant. Calculation of true glottal flow and its components. *Speech Communications*, pages 167-184, 1982.

[6] L.R. Yanguas, G.C. O'Leary, and M. A. Zissman. Incorporating linguistic knowledge into automatic dialect identification of Spanish. To appear in Proceedings, ICSLP November-December 1998.

[7] M. A. Zissman, T. P. Gleason, and D. M. Rekart and B. L. Losiewicz. Automatic dialect identification of extemporaneous, conversational, Latin American Spanish speech. In Proceedings ICASSP, volume 2, pages 777-780, May 1996.

[8] L.R. Yanguas and K.Berkling. Toward further automating dialect discrimination in Spanish. Submitted to Eurospeech, September 1999.