University of Pennsylvania Working Papers in Linguistics

Volume 17
Issue 2 *Selected Papers from NWAV 39*

Article 7

1-1-2011

For the Record: Which Digital Media Can be Used for Sociophonetic Analysis?

Paul De Decker
Memorial University of Newfoundland, pauldd@mun.ca

Jennifer Nycz
Reed College, jnycz@reed.edu

For the Record: Which Digital Media Can be Used for Sociophonetic Analysis?

Abstract
Sociolinguists now have more options for collecting speech data than ever before. Informants who might otherwise be inaccessible to the analyst could record themselves using smartphones and personal computers; researchers might also consider YouTube or recordings uploaded to other internet sites as sources of data. However, digital compression techniques simplify some acoustic content and discard others (Bulgin, De Decker & Nycz 2010) while lower-quality microphones may distort the quality of the signal (Van Son, R.J.J.H. 2005). Therefore, before such data can be used for dialect research, it is crucial to determine if these media affect the reliability of acoustic analyses (Gonzalez, Cervera and Llau 2003; Gonzalez and Cervera 2001). As an initial test, we looked the effect of these devices on representations of the vowel space. Male and female speakers were recorded reading a word list containing 10 English monophthongs in h_d context using a Roland Edirol R-09 (WAV format) recorder, an Apple iPhone (lossless Apple m4a), a Macbook Pro running Praat 5.1 (WAV) and a Mino Flip video camera (AVI converted to AIFF). The Mino Flip file was then uploaded to Youtube and subsequently downloaded (MP3) for analysis. Speakers read each word 3 times while seated in a quiet room with the recorders placed on a table in front of them. Measurements of F1 through F4 were taken at the temporal midpoint of each vowel using Praat 5.1. Differences between recording formats were tested in R using a Repeated Measures ANOVA with separate runs for each formant (F1-F4). Preliminary results indicate that the Mino and Mino-derived YouTube formats differ substantially from the lossless Edirol recording. F1 values for most vowels were raised in Mino and Youtube measurements. F2 was also affected, such that front vowels were artificially raised while back vowels were lowered. Thus the vowel space is effectively altered with lowering along the F1 dimension and a widening of the space along the F2 dimension. These effects seem to be exaggerated for the female speaker. Based on these results, Macbook Pro and iPhone may be suitable recording options for studying the vowel spaces of speakers. Mino and its Youtube derivative show a number of significant deviations from lossless recordings indicating that audio from these devices should not be used for this type of analysis until corrective measures are identified.

For the Record: Which Digital Media Can be Used for Sociophonetic Analysis?

Paul De Decker and Jennifer Nycz

1 Overview

Linguists now have more options for collecting naturalistic speech data than ever before. For example, remote informants might record themselves and other speakers using smartphones and personal computers, enabling scholars to cheaply and quickly gather large amounts of data from far-flung locations. Researchers might also consider drawing on the vast reserves of recorded speech freely available on YouTube or other internet sites as sociolinguistic data sources.

Before such data can be used for sociophonetic research, however, it is crucial to determine the degree to which these media affect the reliability of acoustic analysis (Gonzalez and Cervera 2001, Gonzalez et al. 2003). The digital compression algorithms used in commercially available recording devices simplify some frequency content and discard others (Bulgin et al. 2010), while lower-quality microphones may also distort the acoustic signal. In this paper we compare and evaluate three widely available recording devices (an Apple iPhone, a Macbook Pro, and a Mino (Flip) Video Recorder) for their ability to faithfully reproduce the spectral properties associated with vowel productions recorded to a lossless digital audio recorder that might be currently used in the field (Edirol).

2 Background

2.1 Digital Recoding Devices are Everywhere

Currently, recordings for linguistic analyses are typically made on high quality yet costly machines, which places them out of the hands of most people. While it is unreasonable to expect our informants to have access to the same technology we have in our labs, they may in fact already own any number of other recording devices that could be used for recording linguistic data. Desktop and laptop computers are commonplace in many communities and mobile devices are becoming increasingly common. For instance, a Morgan Stanley report (Yudu Media 2010) indicates that 3.3 million iPads were sold in the first three months of release in 2010; by the end of fiscal year 2010, 73.5 million units were sold worldwide (Kumparek 2010). As of March 2011, 108 million iPhones, 19 million iPads, and 60 million iPod touch units were in use (Dilger 2011). The ubiquity of these devices means the potential for user-generated recordings is high. These recordings could be transmitted to researchers via CD/DVD or via the Internet for use in sociophonetic studies. In addition, researchers might look to tap the seemingly endless hours of audio recordings already available online.

2.2 Youtube and Other Websites

Since its first video upload in 2005, the website Youtube has become hugely popular, with nearly 700 billion "playbacks" initiated in 2010. According to Youtube, "13 million hours of video were uploaded during 2010 and 35 hours of video are uploaded every minute," by speakers ranging in age from 18-54 (YouTube 2011). While not all of these 13 million hours contain speech, the sheer volume of uploads demonstrates that people are familiar with the site and capable of using it as a tool for disseminating audio-video recordings.

Users have already recognized the importance of Youtube for posting instructional or "how-to" videos (approximately 1,830,000 results as of June 1, 2011). Some of these relate to language, dialect and accent demonstrations; a keyword search on 'Accent' yields approximately 79,900 results. These videos feature ordinary people sitting in front of their webcams, showcasing their native dialect. The content of these recordings are often not unlike those found in some sociolinguistic

recording tasks. The end result, in many cases, is a relatively clear video recording of a user's dialect. While Youtube is probably the best-known site of this kind, there are many other websites which feature user-generated content of potential linguistic interest.

While the recording devices and websites discussed here have the potential to yield vast amounts of speech data, they must be approached with caution: most use compression algorithms to reduce file size, and compression may affect the spectral properties that are important for phonetic analysis.

2.3 Compression

Several studies have looked at the effect of compression on spectral properties. For instance, Van Son (2005) found compression algorithms introduced "jump errors" in the range of 3% affecting vowel pitch and formant measurements, suggesting that "only small systematic effects on measurements were found that could be attributed to compression." A similar negligible effect was found by Bulgin et al. (2010), who observed that measurement data from MP3 compression were not significantly different from those taken from uncompressed (WAV) recordings. On the other hand, Bulgin et al. (2010) also found that the compression used by Skype seriously altered vowel measurements, and Rozborski (2007) notes that compression at any level distorts the signal to some degree, with higher rates leading to significant destruction. What we would like to know is the extent to which these findings extend to the digital recording devices and websites that have become popular in recent years and whether or not they are suitable as sociophonetic data collections tools.

3 Data Collection

Speech data was collected from two speakers, one male and one female. The male speaker is the first author, a native of Ontario, Canada; the other speaker is a 23 year old female from Nova Scotia, Canada who was completing her undergraduate degree at Memorial University of Newfoundland at the time of recording. Each speaker was seated, separately, in a sound-attenuated lounge at the Memorial University Sociolinguistics Laboratory (MUSL) and read aloud a word list containing ten English monophthongs in the context [h_d] (Table 1); the word list was read three times, yielding a total of thirty vowel tokens per speaker.

heed	hayed	had	hoed	who'd
hid	head	hod	hud	hood

Table 1: The word list.

Following Byrne and Foulkes (2004), speakers were recorded to four devices simultaneously. All recorders were placed on a table in front of the speaker as he or she read from the Word List. Efforts were made to position the devices at roughly equal distances from the speaker (approximately 30 inches). The first device was a Roland Edirol R-09 recorder, using its built-in stereo condenser microphone. This produced an uncompressed WAV file (44,1kHz sampling rate with 24 bit resolution) used as a baseline for comparison with the other devices. The second device, an Apple first generation iPhone running a proprietary "app," Voice Memo, produced a lossless m4a file format. Using the recording function in Praat 5.1 on a Macbook Pro with a built in microphone, a third uncompressed WAV file was created. The fourth recording, made to a Mino (Flip) Video recorder, produced an MPEG-4 Part 2 (AVI) file. In order to analyze the audio component, an AIFF audio file was extracted from the video file using Apple's iMovie software. Finally, a fifth recording was derived from the Mino (AVI) file, which was uploaded to YouTube and subsequently downloaded in MP3 format and converted to a WAV file for analysis in Praat.

4 Analysis

What follows is an initial test of the devices mentioned above. For some cases (i.e. the Mino video recorder and Mino-derived Youtube recording) we are examining the effects of compression. For

others (i.e. iPhone and Macbook Pro), it is a test of the quality of the microphone.

4.1 Acoustic

All audio files were analyzed using Praat speech analysis software. The vowel portion of each token was segmented, with the beginning of each vowel marked at the onset of a periodic voicing pattern found in the waveform and endpoint marked at the end of periodicity. Measurements of F1 through F4 were taken at the temporal midpoint between these boundaries using the LPC Burg algorithm. Recordings were temporally synced, so that the same timepoint in each utterance was measured across all five recordings of that utterance.

4.2 Statistical

Our statistical analysis of these data aimed to answer two main questions. First, do formant measurements for the same sounds differ significantly depending on the device which records them? Second, if there are differences in measurements across recorders, are these differences greater in certain regions of the vowel space? To this end, a repeated measures ANOVA was run for each of F1, F2, F3, and F4 for each speaker. Recording type (Edirol, iPhone, MacBook, Mino, or Mino-derived YouTube) was included as a within-subjects factor, while Phonological Vowel Height (High, Mid or Low) and Backness (Front or Back) were included as between-subjects factors; each model also included two interaction terms for Type::Height and Type::Backness.

The repeated measures ANOVA only reveals whether there is some difference based on Type, Height/Backness, or an interaction between these factors; it does not reveal precisely where these differences are. Unfortunately, there is no clearly appropriate post hoc pairwise test available for repeated measures ANOVA. Rather than present the results of many (unwieldy and dubious) pairwise comparisons, we will instead present the results of each ANOVA along with visualizations of the data, and discuss where the clearest differences seem to be.

5 Results

5.1 Female Speaker

A significant main effect of Type was found for each of F1 ($F(4,104) = 41.4228$, $p < 0.001$), F2 ($F(4,104) = 3.1635$, $p < 0.01$), F3 ($F(4,104) = 4.2036$, $p < 0.01$), and F4 ($F(4,104) = 12.5294$, $p < 0.001$) for the female speaker.[1] Figure 1 suggests where the significant differences may be between these recordings, at least with respect to F1 and F2. While measurements from each of the four non-Edirol recordings deviate from those of the Edirol to a certain extent, they fall into two groups: the Apple product measurements tend to be similar, while the Mino and Mino-derived YouTube measurements also tend to overlap.

Do the recording-based differences vary depending on where in the vowel space measurements are being taken? The only significant interaction effects were found with F2: both Type::Height ($F(8,104) = 2.0434$, $p < 0.05$) and Type::Backness ($F(4,104) = 6.0487$, $p < 0.001$) emerged as significant. Figure 1 indicates that the Type::Backness result might be due to the Mino and Mino-derived YouTube measurements, which seem to stretch the vowel space along the F2 dimension, resulting in front vowel F2 measurements that are a bit higher, and back vowel F2 measurements that are a bit lower, than those taken from the other devices. The Type::Height interaction, meanwhile, seems to implicate all recordings: while there is an overall effect of increasing F1 measurements across recording types, the difference between types is more extreme for the high vowels.

[1] For each set of analyses, there are uninteresting significant main effects of Phonological Height and Backness on formant realizations: unsurprisingly, Phonological Height is a significant predictor of F1, and Phonological Backness is a significant predictor of F2 and F3 for both speakers.

5.2 Male Speaker

A significant main effect of Type was found for each of F1 (F(4,104)=78.5825, p <0.001), F3 (F(4,104) = 6.3404, $p < 0.001$), and F4 (F(4,104) = 43.0274, $p < 0.001$) for the male speaker. For F1, there were also significant interactions between Type and Height (F(8,104) = 2.7409, $p < 0.001$) and Type and Backness (F(4,104) = 5.4034, $p < 0.001$); figure 2 indicates that these results are likely due to the Mino and Mino-derived F1 values being somewhat higher than those for the Edirol and Apple products, particularly for Low and Back vowels. While there was no significant main effect of Type on F2, there was a significant interaction between Type and Backness (F(4, 104) = 10.8957, $p < 0.001$). In addition to the main effect of Type on F3, there was also a significant interaction between Type and Backness (F(4, 104) = 7.5316, $p < 0.001$). Figure 3a shows that F3 measurements for phonologically Back vowels across devices are more tightly clustered for the male speaker, while the measurements for Front vowels vary (with Mino and Mino-YouTube derived measurements tending to be lower than those taken from other devices). Finally, there was a significant interaction between Type and Backness in the F4 analysis (F(4, 104) = 7.4473, $p < 0.001$).

6 Discussion, Conclusions, and Other Considerations

The results described above, though limited in their scope, indicate that recording device can affect formant measurements. However, the seriousness of this recorder-related variation depends on both the particular device and the use to which its data might be put.

Recordings made with the Macbook Pro and iPhone Voice Memo may be useable for speech analysis, at least of the first and second formants, as the overall shape of the vowel space does not seem to be greatly affected by these devices. However, the greater variation between devices along the F3 dimension (as depicted in figure 3) would mean that any normalization method (such as Bark normalization) that depends heavily on reliable F3 measurements would be ill-advised. Moreover, comparing the speech of speakers who were recorded with different devices would not be recommended.

The Mino recording shows a number of large deviations from lossless recordings indicating that audio from this device and those like it should not be used for sociophonetic analysis. The Mino-derived YouTube recordings shows similar deviations from the lossless recordings, but interestingly, does not differ vary greatly from the original Mino recording in many cases. It is still unclear what this finding means for YouTube-based recordings in general. It may be the case that the Mino recording was already too compressed to be further altered by YouTube; if a lossless file was uploaded to the site, it may show more substantial changes. In future research, we will look at the effects of audio upload on different types of source files, to determine to what extent modifications to such files affect acoustic measurements.

Of course, the few comparisons reported here vastly under-represent the range of acoustic analyses that sociophoneticians might conduct. We have focused on analysis of F1 and F2, and to a lesser extent, F3 and F4.

Beyond the technical issues presented by these and similar devices, there are of course other practical issues surrounding the use of found data or data otherwise collected remotely with speaker-owned recorders. For example, we should want to know if the people who use mobile devices or upload media to sites like YouTube are representative of the larger geographic community under study. When we consider the iPhone alone, 50% of its users are under the age of 30, and 15% are students (Rubicon Consulting Inc. 2008). While this might not be surprising, these data do however point towards a potential disadvantage: if we are interested in a larger representation of people over the age of 30 we are probably not going to find them. Likewise, according to the report, iPhone users tend to be "early adopters of technology" working in "professional and scientific services, arts and entertainment, and the information industry" (ibid.). They are generally not "lower-income workers... late adopters of personal technology" or found in "manufacturing, retail and wholesale sales, health and social care, and food and travel services" (ibid.). Thus, studies which require a wide sampling of different class and age groups may not find iPhone-collected data very convenient or relevant.

With these drawbacks in mind, new technology might prove to be a useful form of data collection. While certain sections of the population might not be represented, researchers who are interested in the population of users that do use these devices might find a wealth of data available to them.

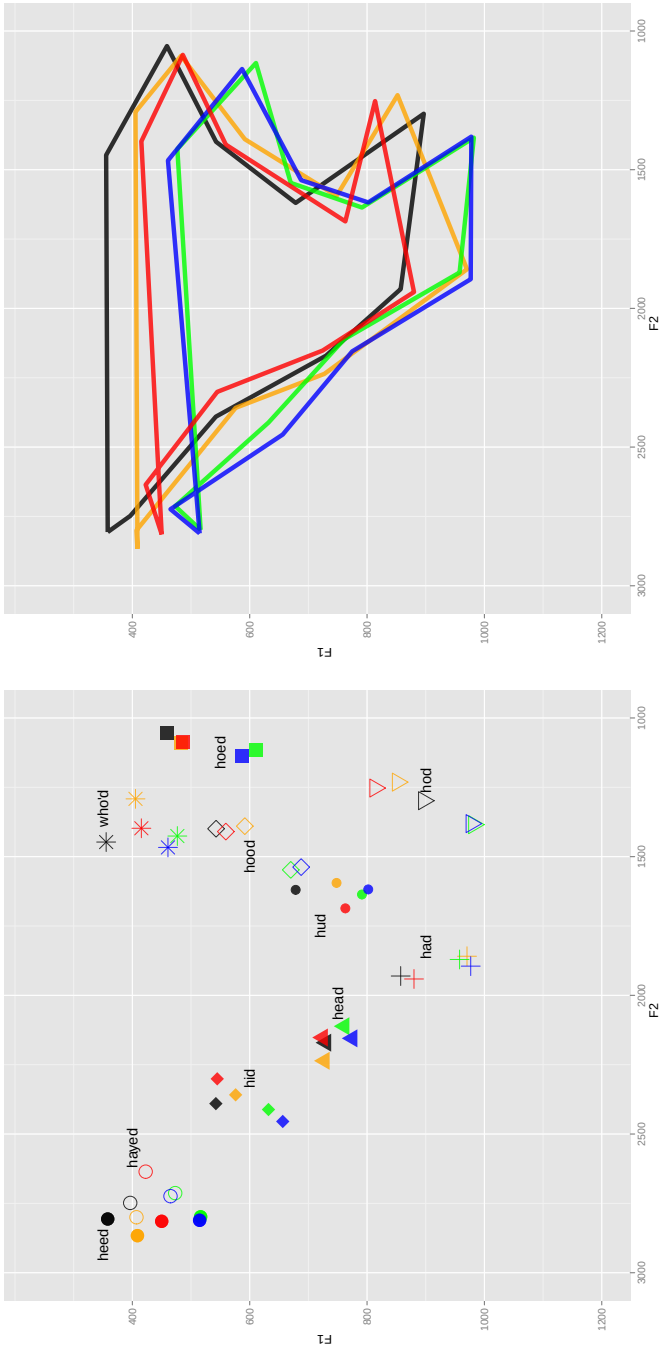


Figure 1: Comparison of F1 and F2 values yielded by each recording type (Female Speaker). 1a plots the mean F1 and F2 values for each Word according to recording Type. 1b connects these points to enable a comparison of the vowel space defined by each recording type.

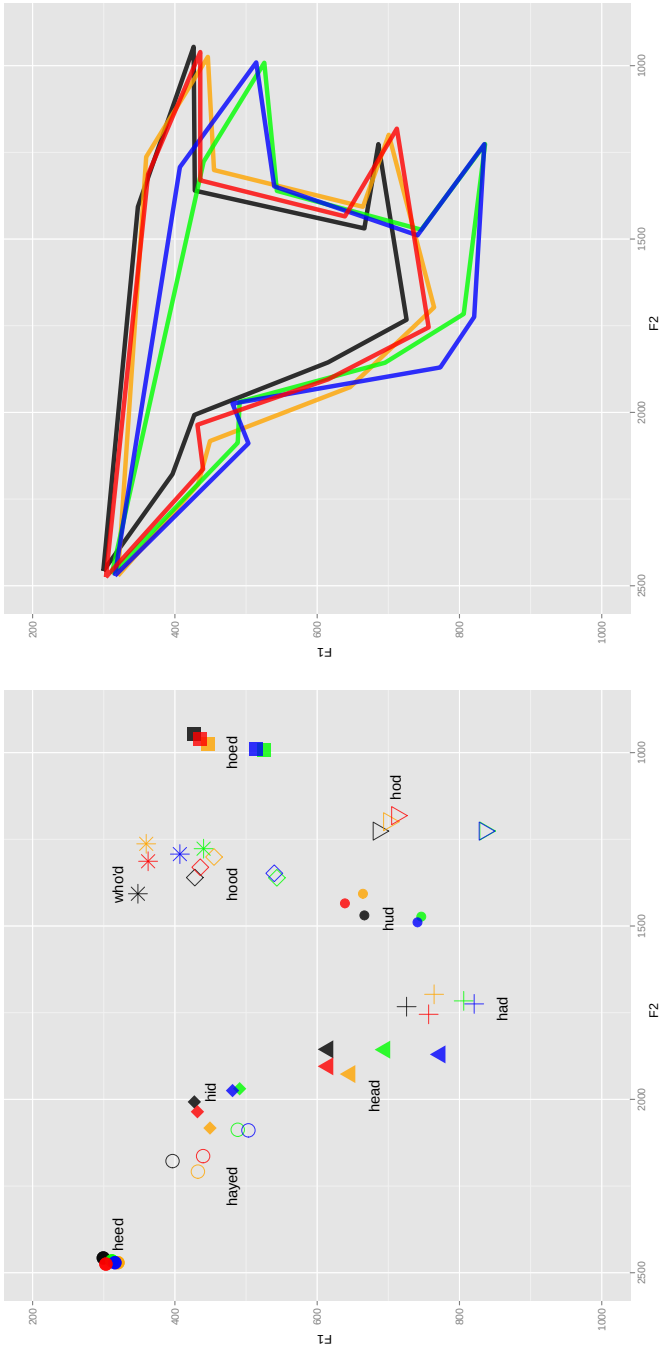
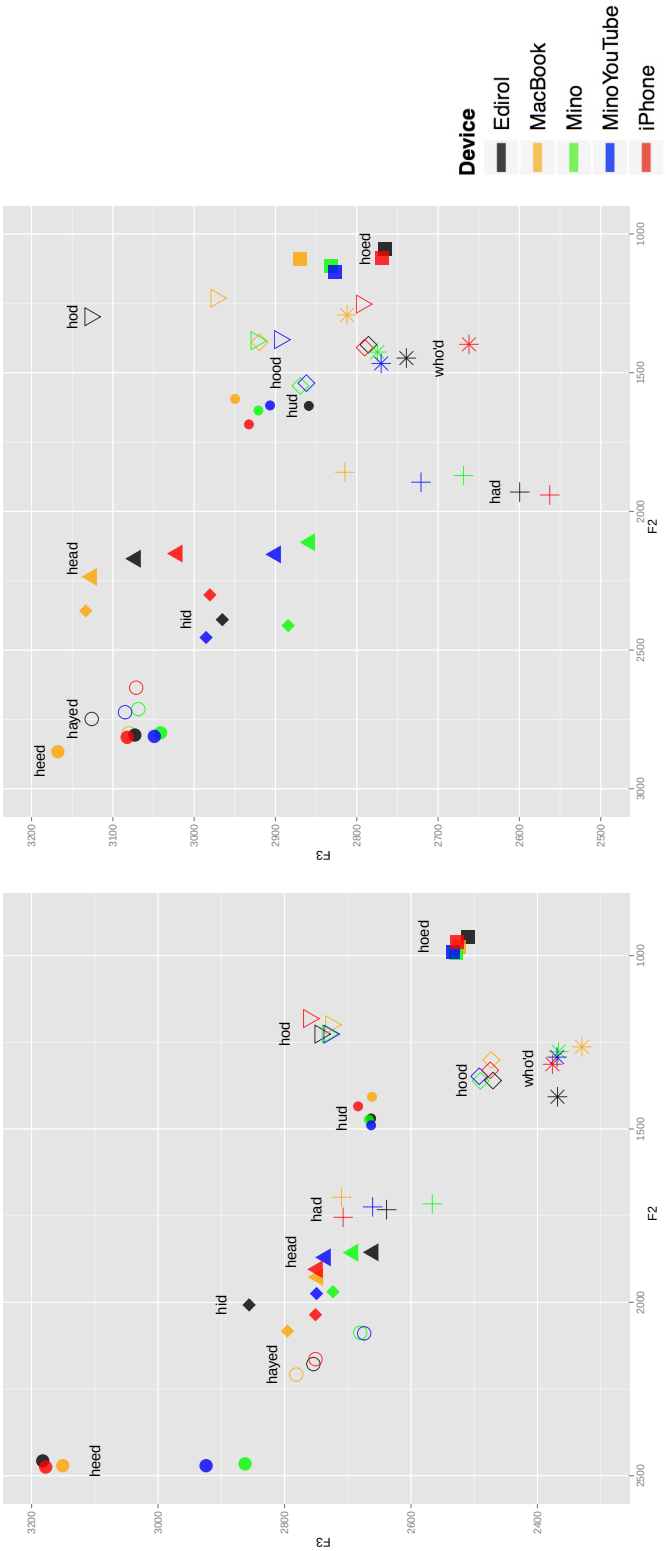


Figure 2: Comparison of F1 and F2 values yielded by each recording type (Male Speaker). 2a plots the mean F1 and F2 values for each Word according to recording Type. 2b connects these points to enable a comparison of the vowel space defined by each recording type.



(a) Male Speaker

(b) Female speaker

Figure 3: The effect of Type on F3 measurements.

References

Bulgin, James, Paul De Decker, and Jennifer Nycz. 2010. Reliability of formant measurements from lossy compressed audio. Poser presented at The British Association of Academic Phoneticians Colloquium, University of Westminster.

Byrne, Catherine, and Paul Foulkes. 2004. The 'mobile phone effect' on vowel formants. *Speech, Language and the Law* 11:83–102.

Dilger, Daniel Eran. 2011. Apple's Samsung lawsuit reveals over 60 million iPod touch sold. Retrieved May 1, 2011, from `http://www.appleinsider.com`.

Gonzalez, Julio, and Teresa Cervera. 2001. The effect of MPEG audio compression on multidimensional set of voice parameters. *Logopedics Phoniatrics Vocology* 26:124–138.

Gonzalez, Julio, Teresa Cervera, and M. Jose Llau. 2003. Acoustic analysis of pathological voices compressed with MPEG system. *Journal of Voice* 17:126–139.

Kumparek, Greg. 2010. Apple sold 14.1 million iPhones last quarter, over 70 million since launch. MobileCrunch. Retrieved May 1, 2011, from `http://www.mobilecrunch.com/2010/10/18/apple-sold-14-1-million-iphones-last-quarter-over-70-million-since-launch/`.

Rozborski, Bogdan. 2007. A preliminary study on the influence of sound data compression upon formant frequency distributions in vowels and their measurement. In *Proceedings of ICPhS XVI, Saarbrucken*.

Rubicon Consulting Inc. 2008. Apple iPhone: Successes and challenges. Retrieved March 31, 2008, from `http://nilofermerchant.com/Rubicon-iPhone_User_Survey.pdf`.

Van Son, Rob J.J.H. 2005. A study of pitch, formant, and spectral estimation errors introduced by three lossy speech compression algorithms. *Acta Acustica united with Acustica* 91:771–778.

YouTube. 2011. Youtube press statistics. Retrieved May 1, 2011 from `http://www.youtube.com/t/press_statistics`.

Yudu Media. 2010. Apple iPad trends and statistics. Retrieved May 1, 2011, from `http://www.doxtop.com/browse/e5ab3c6b/apple-ipad-trends-and-statistics.aspx`.

Paul De Decker
Department of Linguistics
Memorial University of Newfoundland
St. John's, NL A1C 5S7
pauldd@mun.ca

Jennifer Nycz
Department of Linguistics
Reed College
Portland, OR 97202
jnycz@reed.edu