INTERSPEECH 2009 BRIGHTON

Large-Scale Analysis of Formant Frequency Estimation Variability in Conversational Telephone Speech*

Nancy F. Chen [1], *Wade Shen*[1], *Joseph Campbell*[1], *Reva Schwartz*[2]

[1]MIT Lincoln Laboratory, Lexington, MA, USA
[2]United States Secret Service, Washington DC, USA

nancyc@mit.edu, swade@ll.mit.edu, jpc@ll.mit.edu, Reva.Schwartz@usss.dhs.gov

Abstract

We quantify how the telephone channel and regional dialect influence formant estimates extracted from Wavesurfer [1, 2] in spontaneous conversational speech from over 3,600 native American English speakers. To the best of our knowledge, this is the largest scale study on this topic. We found that F1 estimates are higher in cellular channels than those in landline, while F2 in general shows an opposite trend. We also characterized vowel shift trends in northern states in U.S.A. and compared them with the Northern city chain shift (NCCS) [3]. Our analysis is useful in forensic applications where it is important to distinguish between speaker, dialect, and channel characterisitcs.

Index Terms: formant frequency, Northern city chain shift, spontaneous conversational speech, telephone channel, American English, forensic phonetics

10.21437/Interspeech.2009-627

1. Introduction

In this work, we quantitatively investigate how the telephone channel and regional dialect might impact formant frequency estimates extracted from tools commonly used in law enforcement. In 90% of forensic cases, the speech sample in question is recorded after transmission via telephone [4]. Although the band-pass filter characteristics of telephone transmission are well-known, little attention has been paid to its effect on the measurement of acoustic parameters. References [4] and [5] have shown that telephone channels affect estimates of F1 the most, causing an average upward shift of the original frequency for cellular and landline channels of 29% and 5%, respectively. However, many factors that might influence formant estimation variability were not considered in these studies. For example, the number of subjects was limited (at most 20 subjects), only read speech was analyzed (instead of spontaneous conversational speech), and regional dialects were not taken into account either.

Quantifiable norms of dialect-dependent features are necessary for forensic examiners to assess if a given acoustic feature is speaker specific or commonly found in the speaker's dialect [6]. Many researchers have studied the acoustic characteristics of vowels in American English (e.g., [3], [7], [8], [9]). For example, in the Atlas of North American English (ANAE) Labov *et al.* have proposed the Northern city chain shift (NCCS), which is characterized by a clockwise rotation of the low and low-mid vowels: /ae/ is raised and fronted; /eh/, /ah/ and /ih/ are backed; /ao/ is lowered and fronted; /aa/ is fronted [3]. This dialect is referred to as the Inland North dialect of American English, which is spoken in cities along the Erie Canal and in the Great Lakes region, as well as a corridor extending across central Illinois from Chicago to St. Louis [3]. References [8] and [9] have also shown trends that correspond with NCCS using read words. However, the total number of speakers these studies examined are relatively limited (at most 439 speakers collapsed across dialects in [3]).

To fill in the gaps of previous research, we quantify how the telephone channel and regional dialect might influence common algorithms used to estimate formant frequencies in spontaneous conversational telephone speech from more than 3,600 native American English speakers. If these algorithms are to be used in forensic analysis, proper testing should be carried out.

*This work is sponsored by the Command, Control and Interoperability Division (CID), which is housed within the Department of Homeland Security's Science and Technology Directorate under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. Materials and Methods

2.1. Corpus

When estimating formants in spontaneous conversational speech, strong coarticulation and short vowel durations often cause challenges to common formant estimation algorithms [4]. Despite these challenges, we still chose to analyze spontaneous conversational speech over read speech because it is much more realistic in forensic applications. We analyzed more than 1,224 hours of spontaneous conversational speech from 3,673 native American English speakers in the Fisher corpus [10]. These telephone conversations were collected through landline, cordless or cellular phones.

Speakers were categorized according to gender, raised region, and channel. Speakers who grew up in the states where the northern cities are located [3] (New York, Ohio, Wisconsin, Illinois, Indiana, Michigan and Missouri) were considered "northern" speakers, while speakers who grew up in other states were considered as "other". Each regional group contains the same number of tokens from each gender and channel. Similarly, gender and regional factors are equally distributed when examining channel effects. Table 1 lists the number of tokens of each vowel in each channel subcategory: cellular male, cellular female, landline male, landline female. For example, there are 22,478 tokens of /ae/ in each of the 4 subcategories, and the total number of tokens examined for /ae/ is 89,912. The number of tokens in each region subcategory is the same as that in each channel subcategory, thus they are not listed to save space. While the number of tokens is balanced in each subcategory, the number of speakers is not strictly controlled. The speaker number breakdown according to channel and region is listed in Table 2 and Table 3.

Table 1: *Token number of each vowel in the subcategories: cellular male, cellular female, landline male, landline female.*

Vowel	Token number in subcategory	Total token number
/ae/	22,478	89,912
/aa/	4,052	16,208
/ao/	8,658	34,632
/eh/	22,408	89,632
/ah/	21,052	84,208
/ih/	20,338	81,352
/iy/	8,542	34,168
/uw/	14,734	58,936

Table 2: *Number of speakers for each channel category, by vowel. L indicates "landline" and C indicates "cellular".*

Vowel	C Male	C Female	L Male	L Female
/ae/	277	1,030	644	1,458
/aa/	240	929	541	1,388
/ao/	274	1,054	639	1,508
/eh/	278	1,064	648	1,529
/ah/	290	1,107	684	1,583
/ih/	293	1,108	685	1,587
/iy/	93	351	217	498
/uw/	233	917	549	1,329

Table 3: *Number of speakers for each region, by vowel. N indicates "north" and O indicates "other".*

Vowel	N Male	N Female	O Male	O Female
/ae/	348	1,129	573	1,359
/aa/	299	1,037	482	1,280
/ao/	344	1,157	569	1,405
/eh/	349	1,172	577	1,421
/ah/	366	1,219	608	1,471
/ih/	367	1,220	611	1,475
/iy/	121	387	189	462
/uw/	297	1,010	485	1,236

2.2. Phonetic Transcription

Time boundaries of phonetic transcriptions were generated automatically with the following procedures. Word transcriptions labeled by humans were converted to phonetic transcriptions through a standard American English pronunciation dictionary. Then a standard phone recognition system trained at MIT Lincoln Laboratory [11] was used to obtain the time boundaries of the phonetic transcriptions.

2.3. Formant Frequency Estimation

It is known that the popular open source tool Wavesurfer [2], is prone to inaccurate formant estimation when neighboring formant frequencies are close to each other [17]. However, we chose to use Wavesurfer to estimate formant frequencies since it is a common tool used in phonetic analysis.

Since the amount of data we use is at least one order of magnitude larger than most existing studies, it is impractical to manually extract formant frequencies. We automatically extracted F1 and F2 using the Snack Sound Toolkit (*i.e.,* what Wavesurfer uses) using the default parameters: 12th order linear prediction over a 24.9ms window with a 10ms frame interval [1] [2]. F1 and F2 estimates of the median time-point of each vowel token were extracted.

2.4. Vocal Tract Length Normalization

To compare formant frequencies appropriately across individuals, it is desirable to normalize out the variability caused by vocal tract length differences. Therefore we describe two vocal tract length normalization techniques used.

2.4.1. Log-mean normalization (LMN)

Log-mean normalization (LMN) is a vocal tract length normalization technique often used by linguists [3]. Let G be the global log mean of all formant frequencies and S be the speaker-specific log mean of all formant frequencies. $F_{i,j,k}$ indicates the ith formant of speaker j's kth token. n_j is the total number of tokens spoken by speaker j, and m is the total number of formant frequencies (*i.e.*, $m = 2$ in this paper).

$$G = \frac{\sum_{i=1}^{m}\sum_j\sum_k \ln(F_{i,j,k})}{m\sum_j n_j} \quad S_j = \frac{\sum_{i=1}^{m}\sum_k \ln(F_{i,j,k})}{m \times n_j} \quad (1)$$

Then for each speaker j, we compute a scaling factor $f_j = \exp(G - S_j)$. Formant frequencies $F_{i,j,k}$ are then scaled by f_j to be normalized to $\hat{F}_{i,j,k}$; *i.e.,*

$$\hat{F}_{i,j,k} = f_j F_{i,j,k}. \quad (2)$$

2.4.2. Maximum Likelihood Vocal Tract Length Normalization (ML-VTLN)

Maximum likelihood vocal tract length normalization (ML-VTLN) implements a per-speaker linear frequency scaling of the speech spectrum [12] and is often a standard procedure in automated speech recognition systems.

Let $O_{\alpha,j}$ be the acoustic observations of speaker j with the frequency scale warped by the factor α, and the reference acoustic model λ, which contains the "average" speaker's spectral characteristics. The warping factor α^* that maximizes the average per-frame log likelihood given the reference acoustic model λ is chosen for speaker j by

$$\alpha^* = argmax_\alpha P(O_{\alpha,j}|\lambda). \quad (3)$$

We swept α from 0.75 to 1.25 with steps of 0.025 to empirically find the optimal α^*. We re-synthesized the frequency-warped speech [13], and extracted normalized formant frequency estimates as in Section 2.3.

2.5. Statistical Tests

2.5.1. Vocal Tract Length Normalization

We found that the distributions of F1 and F2 estimates are not Gaussian through Kolmogorov-Smirnov and Lilliefors tests [14]. Therefore, to compare and decide which vocal tract length normalization method is more appropriate to use, we used permutation tests [15], due to its non-parametric nature and computational advantage to deal with large amounts of data.

2.5.2. Vowel Shifts

ANOVA was used to measure the statistical significances of the measured vowel shifts and post-hoc Tukey tests were used to correct for multiple comparisons [16]. The threshold of statistical significance was set to 0.001.

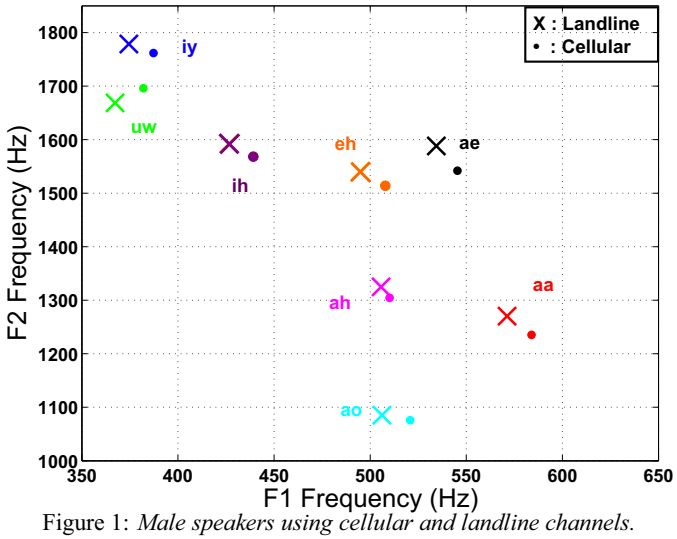

Figure 1: *Male speakers using cellular and landline channels.*

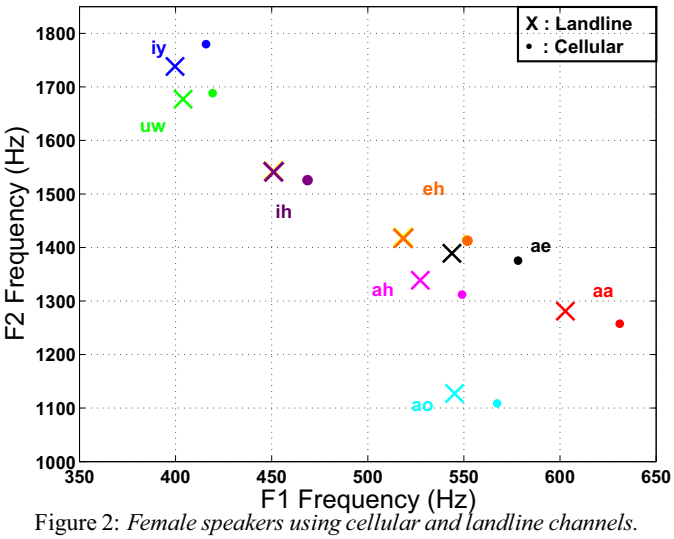

Figure 2: *Female speakers using cellular and landline channels.*

3. Results

3.1. Comparison between LMN and ML-VTLN

To examine how well the normalization methods LMN and ML-VTLN work, for each vowel we performed permutation tests on each channel/region category collapsed across gender. The null hypothesis is that the female and male normalized formant estimates are from the same distribution, which would be what we expect if the formant estimates were normalized appropriately. The p-value represents the probability that the normalized formant estimates of male and female speakers are from the same underlying distribution. We averaged the p-values across channel, region, and vowels. The average p-value of ML-VTLN for both F1 and F2 estimates are larger than those of LMN, indicating that ML-VTLN normalized the raw formant estimates more appropriately. Hence, due to space constraints, we only show results obtained from ML-VTLN.

3.2. Telephone Channel

In Figure 1, we show the means of F1 and F2 estimates produced by male speakers using cellular and landline channels for each vowel. A similar graph for female speakers is shown in Figure 2. Table 4 lists the summary of the statistical tests. These results show that the effect of channel is consistent across vowels and genders for F1; estimates of F1 from cellular channels are higher than those from landline. The trend of F1 corresponds with existing literature: [4] showed that the mobile channel caused F1 estimates to increase an average of 29% when compared to direct recording, while [5] showed that the landline channel caused F1 estimates to increase 5% compared to direct recording. It is thus expected that estimates of F1 from cellular channels to be higher than those from landline channels.

Estimates of F2 for males from cellular channels are generally lower than those from landline (exceptions: /ao/, /uw/), while only half the vowels from females show this trend (/aa/, /ao/, /ah/, /ih/). The lower and upper cut-off frequencies of telephone speech are around 300Hz and 3.2-3.4kHz, which typically are not expected to affect F2 estimates much [4, 5]. However, our results suggest that the telephone channel might actually influence Wavesurfer's F2 estimates.

3.3. Regional Dialect

In Figure 3-4, we show the vowel plot of male and female speakers from northern and other regions. Solid arrows indicate the direction of the NCCS hypothesis. If an arrow is crossed, the vowel does not follow the NCCS hypothesis according to the statistical test, and a dashed arrow is shown to indicate the direction shift if the vowel follows an opposite trend of NCCS. In summary, the following correspond with NCCS: for male speakers, /eh/, /ah/, /ih/ are backed; for female speakers /ae/ is raised and fronted. Vowels showing the opposite trend from NCCS: /ao/ is raised instead of lowered for both genders; /ih/ is fronted for females. The remaining low and low-mid vowels do not show differences between north and other regions.

Table 4: *Statistical test summary for comparing F1 and F2 in different channels. Statistical significance level is set at 0.001. C indicates "cell" and L indicates "landline".*

Vowel	F1 male	F1 female	F2 male	F2 female
/ae/	C>L	C>L	C<L	C=L
/aa/	C>L	C>L	C<L	C<L
/ao/	C>L	C>L	C=L	C<L
/eh/	C>L	C>L	C<L	C=L
/ah/	C>L	C>L	C<L	C<L
/ih/	C>L	C>L	C<L	C<L
/iy/	C>L	C>L	C<L	C>L
/uw/	C>L	C>L	C>L	C=L

The discrepancy between these formant estimate trends and NCCS could potentially be due to the demographic differences between the ANAE and Fisher corpora setup. The sampling strategy for ANAE was to represent the largest possible population, with special attention to those expected to be the most advanced in linguistic changes: speakers were from urbanized area, at least 1 speaker from each city is female between ages 20 to 40, and at least 2-4 speakers were selected in each city [3]. In contrast, the Fisher corpus includes a larger sampling of speakers not limited to urbanized areas, age, or gender. In addition, in this work we explicitly controlled for gender to avoid gender bias. The residential regions in Fisher are documented in states instead of cities, making regional categorization coarser than traditional dialect regions which are linguistically-defined (*e.g.,* ANAE). However, the vowel shift analyses in this work might reveal a more general trend of how typical speakers in northern states sound differently from other speakers in U.S.A., which could be helpful in forensic scenarios when limited demographic information is available.

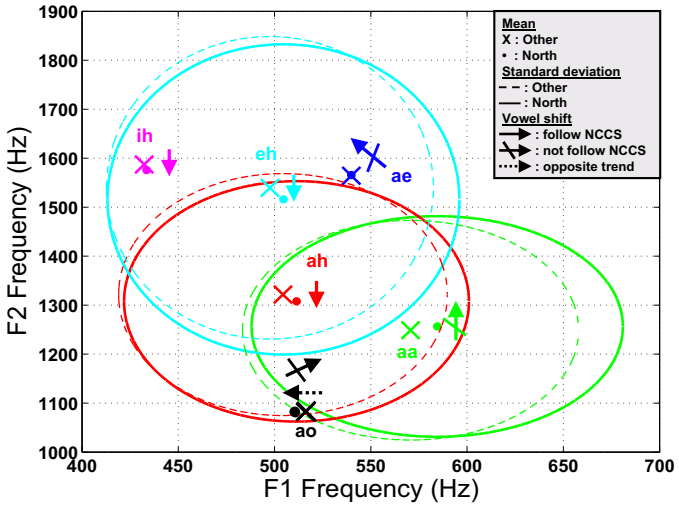

Figure 3: *Male speakers from north and other regions. Standard deviations of some vowels are shown for illustration purposes.*

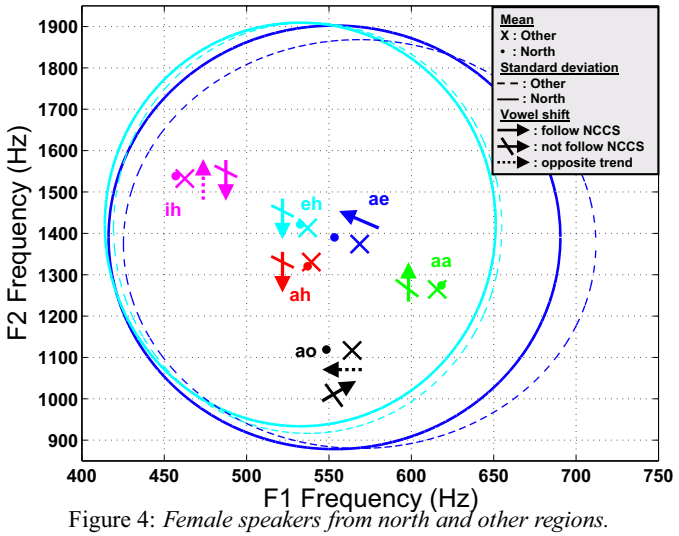

Figure 4: *Female speakers from north and other regions.*

4. Discussion

We should be careful when interpreting the results from Section 3.2 and Section 3.3. These results are effects of channel or region combined with potential formant estimation issues. Strong coarticulation and short vowel durations can cause challenges for common formant estimation tools [4]. In addition, Wavesurfer is especially prone to inaccurate formant estimation when neighboring formants are close to each other [17]. For example, F2 for /uw/ is often overestimated because F2 is low and close to F1. Similarly, F2 and/or F3 for high front vowels, (e.g., /iy/ or /ih/), could be erroneous when close to each other. [17] showed that Wavesurfer's formant estimates are 70 Hz and 94 Hz off for F1 and F2 of vowels using the VTR (vocal tract resonance) database, which is read speech. The high estimation errors of F2 might be one of the potential causes of the observed F2 trend in Section 3.2.

Ideally, if we have speech collected simultaneously through multiple channels (*e.g.,* direct microphones, landline and cellular channels) in parallel, we would be able to tease out the estimation variability caused by the telephone channel and that caused by formant extraction algorithms. Similarly, if corpora documented with linguistically-defined dialect regions are available, we can also further determine if formant estimation tools complicated results in Section 3.3.

5. Conclusion

We quantify how the telephone channel and regional dialect might influence tools commonly used to estimate formants in spontaneous conversational speech from over 3,600 native American English speakers. To the best of our knowledge, this is the largest scale study on this topic. We found that estimates of F1 are higher in cellular channels than those in landline, while F2 in general shows an opposite trend. We also showed the formant estimate trends in northern states of U.S.A. and compared them with NCCS. Our analysis is useful in forensic applications where it is important to distinguish between speaker, dialect, and channel characteristics. Forensic phoneticians should take caution when using common formant estimation tools to analyze formants of telephone speech.

ACKNOWLEDGMENTS

The authors appreciate feedback from Keelan Evanini.

6. References

[1] Snack Sound Toolkit: http://www.speech.kth.se/snack/
[2] Talkin, D., "Speech Formant Trajectory Estimation using Dynamic Programming with Modulated Transition Costs", J. Acoust. Soc. Am., S1, 1987, pp. S55.
[3] Labov, W., Ash, S., and Boberg, C.,"The Atlas of North American English: Phonetics, Phonology, and Sound Change", Mouton de Gruyter, Berlin, 2006.
[4] Byrne, C. and Foulkes, P. "The 'Mobile Phone Effect' on Vowel Formants", Speech, Language, and the Law 11(1), 2004.
[5] Kunzel, H., "Beware of the 'Telephone Effect': the Influence of Telephone Transmission on the Measurements of Formant Frequencies", Forensic Linguistics 8(1), 2001.
[6] Rose, P., "Forensic Speaker Identification", London; New York: Taylor & Francis, 2002.
[7] Peterson, G. and Barney, H. "Control Methods used in a Study of the Vowels," J. Acoust. Soc. Am. 24, 175-184.
[8] Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. "Acoustic Characteristics of American English vowels", J. Acoust. Soc. Am. 97, 3099-3111, 1995.
[9] Clopper, C. G. and Pisoni, D. B. "Acoustic Characteristics of the Vowel Systems of Six Regional Varieties of American English", JASA 118 (3), 1661-1676, 2005.
[10] Cieri C. *et al.*, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text", LREC, 2004.
[11] Shen, W., Chen, N., Reynolds, D., "Dialect Recognition using Adapted Phonetic Models," Interspeech, Australia, 2008.
[12] Hain, T., Woodland, P. C., Niesler, T. R. and Whittaker, E. W. D., The 1998 HTK System for Transcription of Conversational Telephone Speech, ICASSP, 1999.
[13] Quatieri, T.F. and McAulay, R.J., Shape-Invariant Time-Scale and Pitch Modification of Speech, IEEE TASSP, Vol. 40, No. 3, pp. 497-510, March 1992.
[14] Eadie, W.T., Statistical Methods in Experimental Physics. Amsterdam: North-Holland, 1971.
[15] Golland, P. and Fischl, B. "Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies", Information Processing in Medical Imaging, Springer Berlin / Heidelberg, 2003.
[16] Dowdy, S., Wearden, S., Chiko, D. "Statistics for Research", John Wiley & Sons, New Jersey, 2004.
[17] Deng, L., *et al.*, "A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing", International Conference on Acoustics, Speech, and Signal Processing, pp.369-372, 2006.