Effects of mobile phone transmission on formant measurements: a large-scale examination based on 306 Japanese male speakers

Yuko Kinoshita[1], *Takashi Osanai*[2]

[1] College of Arts and Social Science/Asia and the Pacific, The Australian National University
[2]National Research Institute of Police Science, Japan

yuko.kinoshita@anu.edu.au; osanai@nrips.go.jp;

Abstract

This study presents a large scale, well-controlled examination of the effects of mobile phone transmission on the first four formants. We used 306 Japanese male speakers recorded simultaneously with a direct microphone and via a mobile phone network. We found that all four formants were significantly impacted by the mobile phone transmission. Further, we found the impact of mobile transmission was largely unpredictable, and the impacts appear to vary speaker to speaker. This may have significant implications in some application areas, such as forensic phonetics, and also for data collection of speech recorded over phones in general.

Index Terms: formants, mobile phone transmission, Japanese vowels

1. Introduction

The impact of phone transmission on acoustic signals has been of interest in the field of forensic voice comparison (FVC) (e.g. for landlines [1-4], for mobile phone [5-7], and the resulting impact on FVC performance [8, 9]).

For phoneticians, formants have been one of the key acoustic features: the 2011 international survey on forensic voice comparison practitioners (across 15 countries, 36 respondents) reported that 97% conducted some form of formant analysis [10]. While the field is increasingly shifting to techniques based on automatic speaker recognition [11-17], the capacity of formants to link acoustic information to vocal tract configurations is still attractive, as it enables incorporation of linguistic knowledge into interpretation and analysis. It is impossible to gain a full understanding of how, and to what extent, various linguistic and non-linguistic factors affect speech acoustics. However, formants can provide a pathway to meaningful interpretations of a subset of acoustic information.

In this light, we conducted a large scale, well controlled study on the impact of mobile phone transmission (the 'mobile phone effect') on formants. Pre-existing studies are based on populations too small for detailed statistical examination. This study aims to fill this gap by comparing formants extracted from the five Japanese vowels uttered by 306 male speakers of Japanese (40-44 utterances for each vowel) in the NRIPS database [18]. They were simultaneously recorded through two different conditions: a direct microphone; and transmission through a Japanese mobile phone network, which makes this database ideal for our purpose. We extracted the first four formants (F1, F2, F3, and F4), using a semi-automatic approach. We developed this approach to handle the scale of this study: it simulates a decision-making process which a human formant measurer would take in formant selection and corrections, as described in the methodology section.

In this study, we have three objectives. The first is to produce a large formant dataset on two recording conditions, as we believe such data are by themselves useful information.

The second is to examine tendencies in the mobile effect on formants. A previous study on English speakers (six male and six female) compared direct microphone recordings and those transmitted via a mobile network. It reports some concerning results: mobile transmission can significantly shift F1 upwards, especially with high vowels. A majority of speakers had a 20-30% rise in F1, and one male and one female had a 50% and a 40% rise [5]. They found some impact on F2 and F3, though weaker than on F1. They also found that the severity of the mobile transmission effect was not consistent across speakers. Another relevant study examined how the Adaptive Multi-Rate (AMR) codecs of mobile phones affected formants, using three female and five male Australian speakers. It found quite large effects [7], which depended on the choice of codecs, and the effect was greater with higher formants, unlike what was reported in [5]. The female speakers appeared particularly susceptible to this effect, especially with F2 and F3 – the distributions of the formants processed with AMR codecs were far removed from those from the microphone recording. In another study, spectrogram observations showed that all codecs introduce an area of missing energy in the low frequency regions [6]. They pointed out that these white island effects would influence FFT and LPC analyses, which is concerning as formant detection is based on LPC analyses. The characteristics of mobile phone transmission effects change dynamically in response to network conditions; AMR dynamically switches among one of eight sub-codecs in response to its assessment of the condition of its transmission channel. In forensic casework, however, these conditions are unknown to analysts. The large dataset for this study will enable us to explore overarching tendencies at the endpoint of mobile transmission more reliably than these previous small-scale studies, leading to better informed interpretation of forensic casework speech data.

These observations lead us to our third objective: examining whether the severity of the mobile phone effects is speaker dependent. Although we tend to classify speakers into male and female binary sex categories, in reality there are speakers whose speech possesses acoustic characteristics atypical of their biological sex. Considering how severely the codec affected formants of female speakers, we can reasonably speculate that codecs can impact some speakers more severely, even among an all-male population: perhaps those with physically smaller vocal apparatus. If there is a large variation between speakers in how mobile phone transmission affects formants, it calls for careful consideration of what we mean by "comparable" in conducting linguistic analyses. Is matching the conditions of the recording channel and sex of the speaker sufficient, or do we need to control for other factors which contribute to the severity of the impact? This question can have serious implications in some contexts, such as FVC.

2. Methodology

2.1. Database, speakers, and speech materials

This study used 306 male native speakers of Japanese from the NRIPS database [18]. At the time of recording, they were aged 18–76 years and lived in Tokyo and its vicinity, but with varying native dialectal backgrounds. We chose this database for its availability. In the 15 years since it was created, the relevant technologies have evolved greatly, so our results are not directly applicable to the most recent mobile recordings. However, it is still useful data for exploration, and it can provide useful information for analysts, in casework where old recordings needed to be revisited.

All speakers in the NRIPS database were recorded in two non-contemporaneous recording sessions, two to three months apart. They performed the same recording tasks twice at each recording session, and the whole process was recorded simultaneously through multiple channel settings. From these, we chose direct microphone recordings (labelled Ch1) and the same utterances recorded at the receiving end of a mobile phone network (labelled Ch3). The mobile phone used for the data collection was DOCOMO FOMA NEC N902i and the transmission system was W-CDMA. The database was recorded at a sampling frequency 44.1kHz originally, but was down sampled to 8kHz, as Ch3 does not contain acoustic information beyond this.

We selected read-out (C)V syllables as the target speech material, prioritising reliable extraction of formants. The selected consonantal environments are: ∅ (no consonant), /k/, /s/, /t/, /h/, /r/, /g/, /z/, /d/, /b/, and /p/. They were followed by one of the five vowel phonemes of Japanese, /a/, /e/, /i/, /o/, and /u/. This resulted in 11 different phonological contexts for /a/, /e/, and /o/ and ten for /i/ and /u/, as the pairs of syllabary, ぢ /di/ – じ /zi/ and づ /du/ – ず /zu/, are phonetically merged to [dʑi] and [dzɯ] respectively. Every speaker had exactly the same syllables to read out, guaranteeing equal phonological conditions across speakers.

2.2. Formant extraction process

Manual formant measurement is time-consuming and prone to measurer-dependent variability [19]. However, the alternative – automatic extraction – is known to miss or misidentify target formants. Some areas of phonetics have embraced automatic formant extraction, and some new approaches to improve this have been proposed (e.g. [20]). However, they mostly focus on F1 and F2, presumably for the interest in linguistic information rather than speaker information. F1 and F2 have relatively strong spectral energy and are easier to detect. F3 and F4, where more speaker information lies [21-28], are harder to extract automatically, as these spectral peaks are less salient. Unreliable formant detection requires human interventions and corrections, which could introduce supervisor-dependent measurement variability [29], as well as being time consuming. To overcome this, we developed a systematic and replicable approach, which simulates the process that human measurers would apply, as described below.

Step 1: Human measurers often try to measure formants by identifying a stable section of formant trajectory as the target section. Thus, we postulated that the formants of monophthongs can be reasonably represented by the most typical values of the multiple measurements sampled across the duration of a given vowel. We set the formant analysis range of Praat at 0-4kHz and sampled poles every 0.005 seconds. Although our targets were F1 to F4, we chose to extract five poles, since removing peaks that are not our interest is much easier than systematically searching for missed peaks. We called these poles 'peak1-5', not formants, as they may not actually be the target formants.

Step 2: Human measurers can easily detect and reject outliers based on the continuity of the trajectory. We simulated this by producing a histogram with 100Hz bins for peak 1 measurements and identifying the most populated bin and its immediately adjacent bins on both sides. The frequency range represented by these selected three bins was identified as a likely F1 range for this token. All peak 1 measurements which fell in this range were labelled as the updated peak 1. We processed peak 2 in almost the same way: we gathered peak 2 measurements that were higher than the F1 range. We produced a histogram from them, and identified the potential F2 range in the same way as we did with the F1 range. All measurements within the potential F2 range were classified as updated peak 2. For the tokens which did not have a peak 1 measurement that fitted in the likely F1 range, we checked if its peak 2 measurement did. If so, it was considered as a misidentified case and added to the updated peak 1 data. We applied the same process to the measurements from peaks 3 to 5, except the potential frequency ranges were set wider to reflect their naturally greater variations. We identified the most populated bin in the histogram, and the two adjacent bins on both sides of it were used as the potential frequency range. This resulted in time series lists of the measurements of the updated peaks 1-5, which are located within the likely formant ranges. There were a few situations where the histogram appeared bimodal. In such cases, we applied the same process to both peaks and recorded both, giving up to six peaks.

Step 3: We fitted a kernel density estimation to the lists of updated peak measurements, using the density function of the statistical package R. The density function of R automatically assigns x-coordinates by dividing the distance between the minimum and the maximum values in equidistance (we used default 1024 coordinates). The coordinates of the maximum point and its immediately adjacent points are fitted to a quadratic function, and the maximum value of this function was recorded as the most representative frequency value of the formant of a given vowel.

Step 4: From these five (or six, where the token had an additional peak) values, we selected four that are most likely to be F1, F2, F3 and F4. Human measurers would use their phonetic knowledge of likely frequency ranges for each vowel and formant. To simulate this, we firstly calculated the overall distributions of the five peaks for each vowel from all tokens from all 306 speakers (13,464 tokens for /a/, /e/, and /o/ and 12,240 tokens for /i/ and /u/) based on Ch1 (microphone) data, as then are likely to be less affected by external factors. Then, the most typical values for the first four peaks were set as the initial values for F1, F2, F3 and F4.

Step 5: To assign five peaks to four formants, we have five possible patterns (see Table 1. Note that we had 15 patterns, where six peaks were detected). For each token, we selected the pattern in which the four peaks showed the least distance to each of the initial formant values. After performing this process to all tokens, the population means were calculated, and the initial values were replaced with these temporary mean values. This peak assignment process was repeated to re-examine which of the five patterns are the closest to the model population formant values. This process was repeated until the mean formant values stabilised.

As well as formants, we sampled F0 using Praat with the analysis range set at 75-350Hz, as we postulated that F0 may be used as a predictor for the severity of the mobile transmission effect. F0 information was included in the statistical modelling.

Table 1. Possible mapping of peak to formants (for the case of five peaks)

	peak1	peak 2	peak 3	peak 4	peak 5
pattern 1	F1	F2	F3	F4	
pattern 2	F1	F2	F3		F4
pattern 3	F1	F2		F3	F4
pattern 4	F1		F2	F3	F4
pattern 5		F1	F2	F3	F4

2.3. Statistical Analysis

The formant data was first examined with descriptive statistics and visualization, and compared to the results of the previous studies. Then, the mobile phone effect was analysed with linear mixed effects modeling, since our data has a large number of recordings from each speaker, which makes each data point not truly independent [30]. Since the effect of mobile transmission is expected to be non-uniform across the frequency range [6], we built separate models for each formant. We set channel difference, F0, and vowel as the fixed effects, and speakers as a random effect.

3. Results

3.1. Descriptive statistics

3.1.1. *Overall effect*

Table 2 presents the summary statistics from 306 speakers, as well as proportional between-channel differences (Ch3-Ch1 divided by Ch1) in the column 'Diff%'.

Table 2. Summary descriptive statistics

		Ch1		Ch3		Ch3-Ch1
		Mean	SD	Mean	SD	Diff %
a	F1	693	84.4	672	57.8	-3.0%
	F2	1259	92.3	1279	132.9	1.6%
	F3	2632	159.7	2553	131.3	-3.0%
	F4	3516	145.6	3268	178.9	-7.0%
e	F1	494	39.7	486	41.6	-1.7%
	F2	1928	127.1	1914	125.6	-0.7%
	F3	2589	142.2	2519	131.8	-2.7%
	F4	3476	139.1	3211	162.6	-7.6%
i	F1	377	36.5	363	36.5	-3.6%
	F2	2154	159.2	2135	135.9	-0.9%
	F3	2939	139.2	2895	161.4	-1.5%
	F4	3426	111.3	3268	112.5	-4.6%
o	F1	488	48.7	492	44.5	0.8%
	F2	911	98.6	882	70.8	-3.1%
	F3	2689	157.2	2612	172.6	-2.8%
	F4	3357	147.2	3256	102.3	-3.0%
u	F1	394	35.0	398	31.4	0.9%
	F2	1392	150.9	1407	143.2	1.1%
	F3	2380	174.3	2385	128.3	0.2%
	F4	3439	150.6	3287	102.9	-4.4%

In [5], F1 of high vowels in the mobile phone recordings was reported to be 29% higher on average than that of the microphone recordings. Overall, our formant data revealed no such tendency; the mean cross-channel differences for F1 were small, and for /i/ Ch3 were marginally lower. The same study also found that low F2 were lifted and high F2 lowered, but this too was not apparent in our data. For the vowel with the lowest F2, /o/, the mobile phone appears to have had a lowering effect. The vowels with relatively high F2, /i/ and /e/, showed very little difference across the two channels. The only consistent effect of the mobile transmission observed is in F4; the mobile transmission lowers them.

Figure 1 presents the 306 speakers' formants separately for each vowel. Formants extracted from the same utterances but recorded in two different channels (Ch1 and Ch3) were plotted against each other. If the channel difference had no effect on formants, the datapoints should fall very close to the diagonal lines, but that is not the case here. Although the summary statistics did not reveal marked differences between the two channels, the mobile phone transmission appears to have had considerable impact on the formants — and worse still, in mostly unpredictable ways. This suggests difficulties for reliable channel compensation with formants.

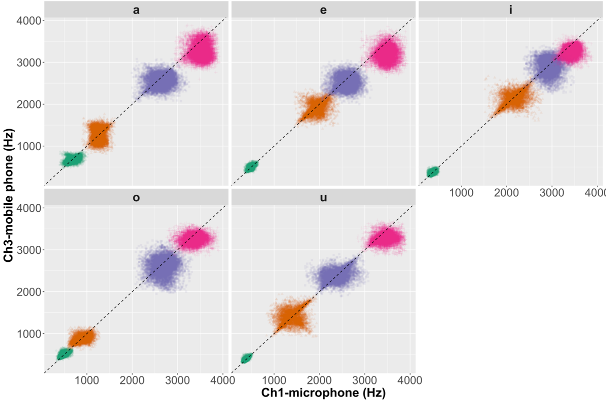

Figure 1. F1- F4 extracted from the same utterances recorded via Ch1 and Ch3 plotted against each other.

3.1.2. *Between-speaker variability of the impact*

Next, we examined how the mobile transmission effect differed between speakers. Figure 2 presents two example speakers, whose mean channel difference was the smallest and the largest among the 306 speakers. Initial visual inspection seems to support our supposition: that mobile transmission does not impact everyone in the same way or to the same extent.

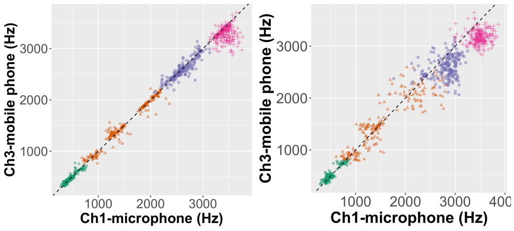

Figure 2. Example of speaker dependency of the mobile effect

3.2. Linear mixed effect model analysis

These initial observations warranted further examination, so we conducted linear mixed effects model analysis. In this section, we examine the between-speaker variability of the mobile transmission effect first, as it determines the better model for overall analysis. We built two models with different slope settings for the random effect (speakers): Model 1 had a fixed slope; Model 2 had a varying slope in relation to the channel effect. Model 2 reflects our hypothesis that the mobile transmission effects on formats are, at least to some degree, speaker dependent. Using the lmer() function of the R lme4 package, each formant was modelled as below:

1. Hz ~ Channel + F0 + Vowel + (1 | Speaker)
2. Hz ~ Channel + F0 + Vowel + (1 + Channel | Speaker)

We then proceeded to test the fit of these models using the R anova() function. Model 2, which allows the slope of the random effect to vary, clearly better performed with all four formants (Table 3). This indicates that different speakers were impacted differently by the mobile transmission.

Table 3. Comparison of the two models.

		AIC	BIC	Chisq	Pr
F1	Model1	1348391	1348479		
	Model2	1346957	1347064	1438.3	< 2e-16
F2	Model1	1605298	1605386		
	Model2	1604810	1604917	492.36	< 2e-16
F3	Model1	1656963	1657051		
	Model2	1654132	1654240	2835.3	< 2e-16
F4	Model1	1642606	1642694		
	Model2	1635885	1635992	6725.4	< 2e-16

Next, we examined the fixed effects based on Model 2: channel, F0, and vowel. Table 4 presents the summary; the columns 'Est', 'SE', 'df', 't', and Pr(>|t|) denote estimate, standard error, degree of freedom, t-value, and p-value, respectively. Our primary interests here are channel effect ('Ch') and F0. 'Int' denotes intercept. We see that all four formants were affected by the mobile phone effect. F0 appeared to strongly predict F1, and less so F2, and had no discernible relationship with F3 and F4. Formants, especially F1 and F2, are closely linked to articulatory gestures. Each vowel has different gestures resulting in different formant frequencies. Vowels as fixed effects are, thus, not of interest in this analysis and are omitted from Table 4. However, we note that the vowel effects were found to be significant for F3 and F4 as well, which was somewhat surprising, as they are generally considered to reflect more speaker information than linguistic information.

Table 4. Fixed effect results

		Est	SE	df	t	Pr(>|t|)
F1	Int	659.9	1.76	1058	374.77	<2e-16
	Ch	-7.3	0.67	305	-10.84	<2e-16
	F0	0.2	0.01	76400	22.19	<2e-16
F2	Int	1278	4.3	1467	297.16	<2e-16
	Ch	-5.61	1.28	305	-4.38	1.67E-05
	F0	-0.05	0.02	52370	-2.03	0.0424
F3	Int	2619	5.27	1225	497.11	<2e-16
	Ch	-53.92	2.8	305	-19.25	<2e-16
	F0	0	0.03	17550	0.17	0.864
F4	Int	3486	4.84	1359	720.17	<2e-16
	Ch	-185.7	3.84	305	-48.33	<2e-16
	F0	-0.01	0.03	25330	-0.42	0.673

Figure 3 presents how speakers' formant values were shifted by transmission through the mobile network. For better visibility, we plotted a randomly selected group of 55 speakers, not 306. Each line represents a different speaker. It shows considerable speaker variations in both the direction and the extent of the impact, suggesting that the overall tendency across the population would not serve well in predicting the mobile transmission effects on an individual's speech recordings. Two possible reasons can be put forward: speakers' spectral characteristics are shaped by their vocal tract configurations, and the impact of switching codecs. However, we speculate that the latter is less likely, as it is reasonable to assume that each speaker's utterance would have been processed with similarly varying codecs.

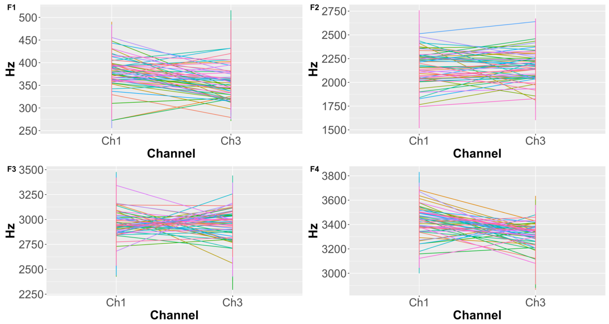

Figure 3. Example of speaker dependency of the mobile phone effect, taken from 55 speakers /i/ vowel.

4. Discussion

These findings point to potentially serious implications. Firstly, use of formants in FVC needs more caution, where it requires comparisons between microphone recordings and recordings made via a mobile phone network. The obtained formants may contain additional and largely unpredictable variability even if the two recordings originated from the same speaker. This increased within-speaker variability makes it less likely for the assessments to result in strong support for a same-speaker hypothesis even when that is factually correct.

Selection of development data for calibration and validation of FVC systems may also require rethinking. A development dataset needs to have 'similar' characteristics to the forensic samples. Gender and channel conditions together with linguistic variety were commonly considered in selecting development data, but our findings suggest this may be inadequate. Further exploration into how we define 'similar' in speech characteristics is in due. Such conditions would also depend on the acoustic features and communication technology in use.

Further, in other types of linguistic research too, phone speech may feature, as it is so ubiquitous. Some caution may be due where this is analysed acoustically.

5. Conclusion

This study performed large scale formant extraction of five vowels from Japanese speech, using an original semi-automatic approach. Through a well-controlled experiment, it revealed that mobile phone transmission affected the first four formants. Closer analysis uncovered a far more complex problem: the mobile phone effect varied across speakers, and its size and direction are difficult to predict. This suggests that the commonly used categories—gender and channel conditions—are inadequate for selecting development data in FVC and further research is needed.

6. Acknowledgement

We thank our anonymous reviewers for their insightful and helpful comments.

7. References

[1] A. Hirson, P. French, and D. Howard, "Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics," in *Studies in General and English Phonetics*, J. W. Lewis Ed. London: Routledge, 1995, pp. 230-240.

[2] H. J. Künzel, "Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies," *Forensic Linguistics* vol. 8, no. 1, pp. 80-99, 2001.

[3] S. Lawrence, F. Nolan, and K. McDougall, "Acoustic and perceptual effects of telephone transmission on vowel quality," *International Journal of Speech, Language & the Law,* vol. 15, no. 2, 2008.

[4] P. J. Rose, D. Lucy, and T. Osanai, "Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical effects model: A "non-idiot's bayes" approach," in *the 10th Australian International Conference on Speech Science & Technology*, Sydney, S. Cassidy, Ed., 8-10/12/2004 2004: Australian Speech Science and Technology Association, pp. 402-407.

[5] C. Byrne and P. Foulkes, "The 'mobile phone effect' on vowel formants," *International Journal of Speech Language and the Law,* vol. 11, no. 1, pp. 83-102, 2004.

[6] B. J. Guillemin and C. Watson, "Impact of the GSM mobile phone network on the speech signal: some preliminary findings," *International Journal of Speech, Language & the Law,* vol. 15, no. 2, 2008.

[7] B. J. Guillemin and C. Watson, "Impact of the GSM AMR speech codec on formant information important to forensic speaker identification," in *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, 2006, pp. 483-488.

[8] A. Alexander, D. Dessimoz, F. Botti, and A. Drygajlo, "Aural and automatic forensic speaker recognition in mismatched conditions," *The International Journal of Speech, Language and the Law,* vol. 12, no. 2, pp. 214-234, 2005.

[9] C. Zhang, G. S. Morrison, E. Enzinger, and F. Ochoa, "Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices," *Speech Communication,* vol. 55, no. 6, pp. 796-813, 7// 2013, doi: http://dx.doi.org/10.1016/j.specom.2013.01.011.

[10] E. Gold and P. French, "An international investigation of forensic speaker comparison practices," in *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China*, 2011, pp. 1254-1257.

[11] A. P. Ajit, A. George, and L. Mary, "I-Vectors for Forensic Automatic Speaker Recognition," in *2018 International CET Conference on Control, Communication, and Computing (IC4)*, 2018: IEEE, pp. 284-287.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018: IEEE, pp. 5329-5333.

[13] M. Jessen, J. Bortlík, P. Schwarz, and Y. A. Solewicz, "Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)," *Speech Communication,* vol. 111, pp. 22-28, 2019/08/01/ 2019, doi: https://doi.org/10.1016/j.specom.2019.05.002.

[14] M. Jessen, G. Meir, and Y. A. Solewicz, "Evaluation of Nuance Forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)," *Speech Communication,* vol. 110, pp. 101-107, 2019/07/01/ 2019, doi: https://doi.org/10.1016/j.specom.2019.04.006.

[15] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, and A. Alexander, "Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)," *Speech Communication,* vol. 112, pp. 30-36, 2019/09/01/ 2019, doi: https://doi.org/10.1016/j.specom.2019.06.005.

[16] G. S. Morrison and E. Enzinger, "Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Conclusion," *Speech Communication,* vol. 112, pp. 37-39, 2019/09/01/ 2019, doi: https://doi.org/10.1016/j.specom.2019.06.007.

[17] J. Rohdin *et al.*, "End-to-end DNN based text-independent speaker recognition for long and short utterances," *Computer Speech & Language,* vol. 59, pp. 22-35, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.csl.2019.06.002.

[18] H. Makinae, T. Osanai, T. Kamada, and M. Tanimoto, "Construction and preliminary analysis of a large-scale bone-conducted speech database," (in Japanese), *IEICE technical report,* vol. Speech 107, no. 165, pp. 97–102, 2007. [Online]. Available: http://ci.nii.ac.jp/naid/40015600747/.

[19] M. Duckworth, K. McDougall, G. de Jong, and L. Shockey, "Improving the consistency of formant measurement," *International Journal of Speech, Language & the Law,* Article vol. 18, no. 1, pp. 35-51, 2011, doi: 10.1558/ijsll.v18i1.35.

[20] K. Evanini, S. Isard, and M. Liberman, "Automatic formant extraction for sociolinguistic analysis of large corpora," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[21] F. Clermont and P. Mokhtari, "Acoustic-articulatory evaluation of the upper vowel-formant region and its presumed speaker-specific potency," in *Fifth International Conference on Spoken Language Processing*, 1998.

[22] S. Furui and M. Akagi, "Perception of voice individuality and physical correlates," *音響学会聴覚研資,* pp. H 85-18, 1985.

[23] U. G. Goldstein, "Speaker‐identifying features based on formant tracks," *The Journal of the Acoustical Society of America,* vol. 59, no. 1, pp. 176-182, 1976.

[24] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication,* vol. 50, no. 4, pp. 312-322, 2008.

[25] P. Mokhtari and F. Clermont, "Contributions of selected spectral regions to vowel classification accuracy," in *Third International Conference on Spoken Language Processing*, 1994.

[26] L. C. Pols, H. R. Tromp, and R. Plomp, "Frequency analysis of Dutch vowels from 50 male speakers," *The journal of the Acoustical Society of America,* vol. 53, no. 4, pp. 1093-1101, 1973.

[27] S. Saito and F. Itakura, "Personal characteristics of the frequency spectrum for vowels," *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics,* vol. 16, pp. 73-79, 1982.

[28] K. N. Stevens, "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds," *Proceedings of the Seventh International Cons. Phonetic Sciences,* pp. 206-232, 1971.

[29] C. Zhang, G. S. Morrison, E. Enzinger, and F. Ochoa, "Laboratory Report: Human-supervised and fully-automatic formant-trajectory measurement for forensic voice comparison–Female voices," *FVC, EE&T, UNSW Laboratory Report,* 2012.

[30] R. Baayen, "Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge University Press," 2008.