THE ROLE OF THE FIRST FIVE FORMANTS IN THREE VOWELS OF MANDARIN FOR FORENSIC VOICE ANALYSIS

Honglin Cao[a,b,c] &Volker Dellwo[c]
[a] Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education, China.
[b] Collaborative Innovation Center of Judicial Civilization, China.
[c] Institute of Computational Linguistics, University of Zurich, Switzerland
caohonglin@cupl.edu.cn, volker.dellwo@uzh.ch

ABSTRACT

Formant characteristics are most commonly part of forensic speaker comparison (FSC). However, only formants F1 to F3 typically occur in evidence material because it is mostly recorded via telephone. Given recent technological advances in telephony (e.g. WeChat or WhatsApp) higher formants (F4-F5) are becoming increasingly part of evidence material. The present study investigated the speaker-distinguishing properties of F1 to F5 of three sustained vowels /i/, /y/ and /ɤ/ in Mandarin produced by 20 young male speakers. Based on discriminant analysis, for each single formant, the best predictors were F5 for /i/ and F4 for /y/ and /ɤ/. Classification performance varied between vowels. Inclusion of two and three formants yielded higher classification rates of 30−80%. The best value was provided by the combination of F2, F4 and F5 of /ɤ/. The value and limitations of F4 and F5 for FSC are discussed.

Keywords: speaker characteristics, vowel, higher formant frequencies, discriminant analysis, Mandarin

1. INTRODUCTION

Formant frequencies are one of the most widely-used parameters in forensic speaker comparison (FSC) [1-5]. Forensic speech evidence is often recorded via telephone/mobile phone, so typically formants F1 to F3 occur in the evidence material. Higher formants usually lie outside the telephone passband (about 300−3500Hz) [6]. According to an international survey conducted by Gold & French [5], a high proportion of 35 expert forensic phonetic analysts reported measuring F1, F2, and F3 (87%, 100%, 87%, respectively) but only 17% obtained measurements of F4 (most likely because of the limited passband). F4 analysis for non-bandlimited speech is more typically obtained (e.g. [7-9]). No studies in FSC could be found that analysed the speaker-distinguishing power of F5.

In the recent past, there were rapid and striking technological innovations in telephony, for example the emergence of systems like WeChat, QQ, WhatsApp and other instant messaging apps. Currently there are increasing numbers of people using these apps for sending audio messages which thus more often appear as evidences in court [10]. These audio messages are characterized by higher quality, mostly in terms of the signal bandwidth (e.g. WeChat uses 8 kHz). This facilitates measurements of higher formants like F4 and F5. Given that these audio messages are often limited in duration (WeChat allows max. 60 sec.) there is a high necessity to make use of any information that is present in the speech signal.

In FSC, it is commonly hypothesised that higher formants carry more speaker-specific characteristics compared to lower ones. The relative degree of speaker-specific information provided by each formant, however, is not clear-cut. Based on the centre measurement of the German vowel /a/, Jessen [11] found that F3 carried more speaker-specific information than F2 and F1. Similarly, Nolan [2] found F3 in English /r/ and /l/ to be more speaker-specific compared to F1 and F2. Using dynamic features of /aɪ/, both McDougall [12] and Hughes [13] found F3 outperformed F1 and F2. When F4 is considered, the picture becomes more complicated. For instance, Rose [7] showed the ranking of the speaker discriminating power of formants in the utterance *hello* was F2, F4, F3 and F1. Based on long-term formant (LTF) distributions (LTF1 to LTF4) of 100 male speakers of English, the results of [9] suggested that LTF3 performed the best overall, followed by LTF4, LTF1, and finally LTF2 in discriminating speakers. In addition, there is a high variability of the speaker discriminating power of formants between different vocalic categories. For example, Kinoshita [8] measured the centre frequencies of F1 to F4 of five Japanese vowels and found all F4 of /a/, /i/, /u/ and /e/ underperformed F3, with the exception of /o/; among all formants, F2 of /e/ was the most promising discriminator, followed by F3 of /e/ and F2 of /i/. Using dynamic properties of F1 to F3 of seven monophthongs in Czech, Fejlov á et al. [14] found that /i:/ and /a:/ outperformed the remaining five vowels. For both /i:/ and /a:/, F2 carried more speaker-specific information than F1 and F3. Morrison [15] compared the dynamic features of F1 to F3 of five diphthongs in Australian English and found the best-performing vowel was /eɪ/

followed by /aɪ/, /oʊ/, /ɔɪ/ and /aʊ/. This demonstrates that speaker-specific information of formants varies for different formants and different vowels.

The present study investigated the relative degree of speaker-specificity of F1 to F5 in three vowels /i/, /y/ and /ɤ/ in Mandarin. Our aims were (a) to assess the speaker discriminating power of different formants in different vowels and (b) to estimate the suitability of F5 measurements for FSC. (a) was addressed by linear discriminant analysis (LDA) and (b) was addressed by analysing the vowels and speakers for which F5 was obtainable.

2. METHOD

2.1. Subjects and Materials

39 male speakers of Chinese aged 19-30 years were recruited. The materials were eight sustained monophthongs /a/, /o/, /ɤ/, /i/, /u/, /y/, /ɿ/ and /ʅ/. /ɿ/ and /ʅ/ are distinctive vowels in Chinese phonology (e.g. in the words "姿/tsɿ55/" and "知/tʂʅ55/"), however, they haven't been accepted as IPA characters yet.

2.2. Recording

A SONY ECM-44B condenser microphone was used to record the materials in a sound-attenuated room at Peking University (sampling rate 22 kHz). In order to simulate the band-pass of audio message via WeChat, all recordings for this study were resampled to 16 kHz. Data was collected at two recording sessions separated by about one week to one month. Within each session, subjects were required to articulate the eight sustained vowels for about one second twice.

2.3. Formant Measurements

Wavesurfer [16] was chosen to extract formant values using a LPC-based algorithm. The steady-state segment of the vowel was chosen by hand for formant tracking. For settings, number of formants and LPC order were adjusted to find the most plausible formant analysis based on visual inspection in a wide-band spectrogram for each particular vowel. The typical setting was 5 formants in 5 kHz signal bandwidth at LPC order 14. For the visual inspection, four displays of each vowel were compared on a computer screen. Ambiguous formant tracks were excluded from the analysis. 1248 vowel samples (39 speakers × 8 vowels × 4 repetitions) were analysed.

2.4. Statistical Analysis

LDA was performed to assess the degree of speaker-specificity of different formants using SPSS 22.0. As a closed-set procedure, the effectiveness of LDA for non-investigative FSC research was demonstrated by a number of studies, e.g. [12, 14, 17-20].

3. RESULTS AND DISCUSSION

We selected each vowel for each speaker, for which F5s of the 4 repetitions were steadily obtainable. For the 8 vowels, F5 was obtainable in the following way: /ɤ/=29, /y/=27, /i/=26, /ɿ/=25, /u/=22, /o/=19, /ʅ/=17 and /a/=14 out of the 39 speakers. This means that in 20 speakers F5 was obtainable in the combination of the top 3 vowels /ɤ/, /y/ and /i/. F1 to F5 of the 3 vowels of the 20 speakers (denoted S1, S2, …S20) were hence selected for further analysis.

Figure 1 shows mean value and ±1 standard deviation (SD) of each formant (F1-F5) for the 4 repetitions of vowel /ɤ/ produced by the 20 speakers. The F4 mean values, which are arbitrarily chosen, are sorted from the lowest to the highest. Differences of formant pattern among speakers were evident: e.g. F2 of S13 and S5 are very similar, while F3, F4 and F5 differ a lot. It can be seen that, for individual formants, between-speaker variation of F4 and F5 of /ɤ/ seem to be larger than that of the other three. But for within-speaker variation, F5 seems to be the largest. When F4 increases across different speakers, only F5 seems to follow generally. In other words, a positive relationship between F4 and F5 of /ɤ/ can be expected.

Figure 1: Line charts for F1 to F5 of the vowel /ɤ/ of 20 speakers. Error bars are given in ±1 SD.

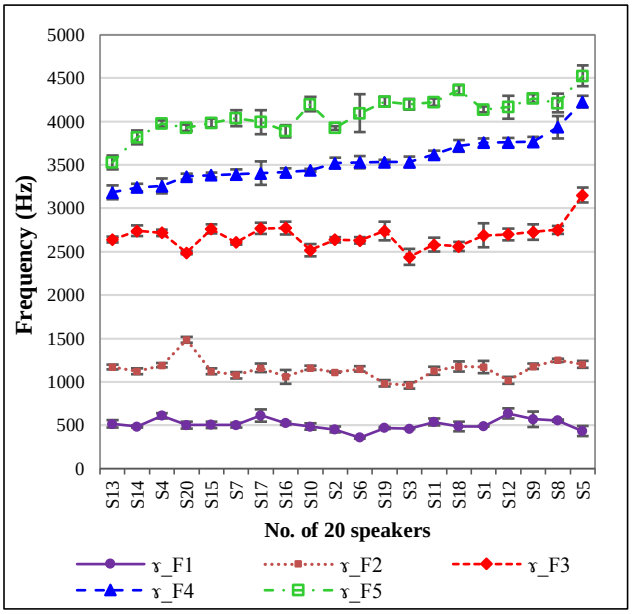
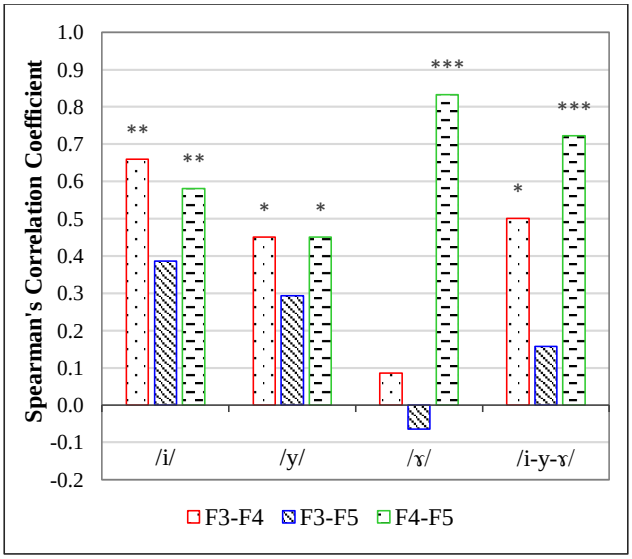


Figure 2 shows the Spearman's correlation coefficients of relationships between the three higher formants, namely F3, F4 and F5, of each vowel as well as the average of the three vowels (/i-y-ɤ/). Consistent with the results displayed in Figure 1, for

vowel /ɤ/, F4 correlated positively with F5 (r= 0.833, p<0.001), but not with F3 (r=0.086). For /i/, the results were slightly different: positive correlations were found for F3 *vs.* F4 (r=0.659, p<0.01), F4 *vs.* F5 (r=0.581, p<0.01), but not for F3 *vs.* F5. The results for /y/ are very similar to those for /i/. Strong correlation was found between F4 and F5 of the averaged across vowels /i-y-ɤ/ (r=0.722, p<0.001) but weak correlation between F3 and F4. For all four conditions, the relationships between F3 and F5 are not significant.

Figure 2: Bar charts for the correlation coefficients between F3, F4 and F5 for /i/, /y/ and /ɤ/ and the average of the three vowels /i-y-ɤ/.
* p<0.05, ** p<0.01, *** p<0.001 (all 2-tailed).



LDA was performed using formant frequencies as predictors of membership of the 20 speakers (S1-S20). Separate analyses were run for each single formant and 20 combinations of 2 or 3 formants of each vowel (the number of formants for combinations are less than 4, because LDA puts a limit to the number of predictors, which should be no more than the number of the tokens [21], namely 4 repetitions in the present study). The classification rates (CR) for each discriminant analysis are summarized in Table 1. In order to display the CR values more clearly, Figure 3 was generated (the CR values of /i/ were sorted from the lowest to the highest). All CR values were greater than chance level (1/20=5%).

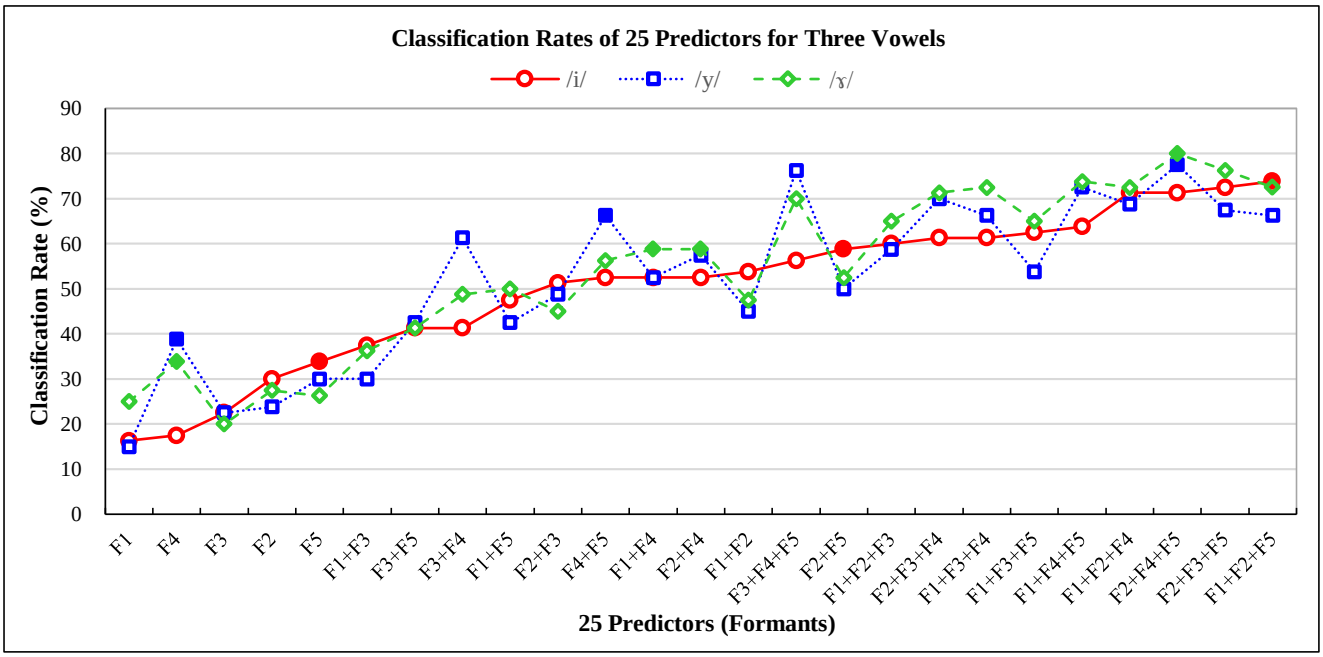
Examining F1-F5 individually, the results suggest that F5 of /i/ (33.8%), F4 of /y/ (38.8%) and F4 of /ɤ/ (33.8%) achieve the highest CRs and the best predictor is F4 of /y/. Specifically, for vowel /i/, F5 performs best, followed by F2, F3, F4, and finally F1, which is in full agreement with the findings of [8] (F5 of /i/ was not analysed in [8]). For vowel /y/, it is unexpected that F4 performs much better than F5 (above 8.8%); F2 underperforms F5, but slightly outperforms F3; F1 achieves the lowest CR (15.0%), which is just half of the CR of F5. For vowel /ɤ/, the CR values were found to be in the following order: F4>F2>F5>F1>F3. Compared to /i/ and /y/, the differentiating values of F1, F2 and F5 of /ɤ/ are very similar. The results suggest that the relative degree of speaker-differentiating value of different formants of different vowels varies markedly, which is consistent with the findings from previous studies [8, 14, 15]. When just F1-F3 are considered, interestingly for all three vowels, it is F2 not F3 that performs the best (cf. [14]). Nevertheless, it is still safe to conclude that generally higher formants tend to convey more speaker discriminant information (e.g. [2, 3, 7, 9, 11-13]). For different vowels, however, the best-performing formant differs.

Table 1: Classification rates (CR) for LDA based on predictors from one, two and three formants of /i/, /y/ and /ɤ/. *NP* means the number of predictors. The largest CR for each subgroup is shown in bold.

NP	Predictors	Classification Rate (%)		
		/i/	/y/	/ɤ/
1	F1	16.3	15.0	25.0
	F2	30.0	23.8	27.5
	F3	22.5	22.5	20.0
	F4	17.5	**38.8**	**33.8**
	F5	**33.8**	30.0	26.3
2	F1+F2	53.8	45.0	47.5
	F1+F3	37.5	30.0	36.3
	F1+F4	52.5	52.5	**58.8**
	F1+F5	47.5	42.5	50.0
	F2+F3	51.3	48.8	45.0
	F2+F4	52.5	57.5	**58.8**
	F2+F5	**58.8**	50.0	52.5
	F3+F4	41.3	61.3	48.8
	F3+F5	41.3	42.5	41.3
	F4+F5	52.5	**66.3**	56.3
3	F1+F2+F3	60.0	58.8	65.0
	F1+F2+F4	71.3	68.8	72.5
	F1+F2+F5	**73.8**	66.3	72.5
	F1+F3+F4	61.3	66.3	72.5
	F1+F3+F5	62.5	53.8	65.0
	F1+F4+F5	63.8	72.5	73.8
	F2+F3+F4	61.3	70.0	71.3
	F2+F3+F5	72.5	67.5	76.3
	F2+F4+F5	71.3	**77.5**	**80.0**
	F3+F4+F5	56.3	76.3	70.0

Results from Table 1 also show that inclusion of 2 and 3 formants yields CR values of 30.0−80.0%. The combination of F2, F4 and F5 of /ɤ/ outperforms all other combination scenarios as well as individual formants. The best combinations for /i/ and /y/ are

Figure 3: Line charts for the classification rates (CR) for LDA based on predictors from one, two and three formants of /i/, /y/ and /ɤ/. The largest CR for each subgroup is shown in solid.



F1+F2+F5 (73.8%) and F2+F4+F5 (77.5%), respectively. These results are also in line with the finding from [9, 12, 15, 17] who found that better classification can be achieved with higher number of predictors. A final LDA was carried out using three of the best predictors of each vowel, namely F5 of /i/, F4 of /y/ and F4 of /ɤ/. The CR value is 80.0%, which indicates that no more improvement is obtained. It is probably because of the high correlations between F5 of /i/ and F4 of /ɤ/ (r=0.605, p<0.01), and between F4 of /y/ and F4 of /ɤ/ (r=0.702, p<0.001) (F5 of /i/ and F4 of /y/, not significant).

A common argument has been made that vowel category information is largely determined by the first two or three formants. By contrast, higher formants (F4, F5, etc.) are always expected to be largely independent of vowel category and carry more speaker individualities, which was replicated in the present study. However, the mechanisms for F4 or F5 are more complicated and relatively little investigated. One possible interpretation is that F4 and F5 are sensitive to the laryngeal cavity (LC) shape (when LC is shortened, F5 and F4 increase) [22]. More recently, Takemoto et al. [23] found that F4 was mainly determined by the LC geometry. Another study conducted by the same research group also found that the shape of the hypopharynx (i.e. laryngeal tube and piriform fossa), regardless of vowel type, showed relatively small within-speaker variation and relatively large between-speaker variation [24], supporting our finding that F4 is one of the best-performing formants.

Our results also showed that, for some vowels of some speakers, F5 cannot be obtained (e.g. F5 cannot be clearly displayed on the spectrogram or be reliably spectrally separated from F4). One plausible reason for this is the strong anti-resonance, caused by piriform fossa, which constantly appears in the frequency region between 4 to 5 kHz (basically that is the region for F5 of adult male speakers) in spontaneously produced and sustained vowels [25].

4. CONCLUSION

Overall, the results presented in this study suggest that the higher formants, F4 and F5, exhibit more speaker-distinguishing power than the lower ones, F1 to F3. The performance of individual formants varies between different vowels. F4, F5 and other acoustic-phonetic features below 8 kHz are worth exploring for FSC purpose based on the increasing WeChat/WhatsApp audio message evidences. It is important to note, however, that not all speakers provide F5 data that is suitable for FSC. In practical terms, it would be desirable to replicate the present results for realistic forensic recording material regarding speaking style and environment.

5. ACKNOWLEDGEMENTS

This research is supported by the Humanity and Social Science Youth Foundation of Ministry of Education of China (No: 18YJC740004) and Program for Young Innovative Research Team in China University of political Science and Law (18CXTD09).

6. REFERENCES

[1] Jessen, M., 2008. Forensic Phonetics. *Lang Linguist Compass* 2(4), 671–711.

[2] Nolan, F., 1983. *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

[3] Rose, P., 2002. *Forensic Speaker Identification*. London and New York: CRC Press.

[4] Foulkes, P., French, P. 2012. Forensic Speaker Comparison: A Linguistic-Acoustic Perspective, In: Solan L.M., Tiersma, P.M. (eds), *The Oxford Handbook of Language and Law.* Oxford University Press: Oxford. 557–573.

[5] Gold, E., French, P. 2011. International Practices in Forensic Speaker Comparison. *Int J Speech Lang La* 18(2), 293–307.

[6] Byrne, C., Foulkes, P. 2004. The 'Mobile Phone Effect' on vowel formants. *Int J Speech Lang La* 11(1), 83–102.

[7] Rose, P. 1999. Differences and distinguishability in the acoustic characteristics of *hello* in voices of similar-sounding speakers: A forensic phonetic investigation. *Australian Review of Applied Linguistics* 22(1), 1–42.

[8] Kinoshita, Y. 2001. *Testing realistic forensic speaker identification in Japanese: A likelihood ratio–based approach using formants*. PhD dissertation, Australian National University.

[9] Gold, E., French, P., Harrison, P. 2013. Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proc. Meet Acoust* 19, 060041.

[10] Cao, H., Ding, T. 2018. An Empirical Study on the Application of Evidence of Forensic Phonetics in Courts of Beijing, Shanghai, Guangzhou, Shenzhen, Tianjin and Chongqing in China. *Evi Sci*. 26(5), 622–638.

[11] Jessen, M. 1997. Speaker-specific information in voice quality parameters. *Forensic Linguistics* 4(1), 84–103.

[12] McDougall, K. 2004. Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *Int J Speech Lang La* 11(1), 103–130.

[13] Hughes, V. 2013. Establishing typicality: A closer look at individual formants. *Proc. Meet Acoust* 19, 060042.

[14] Fejlová, D., Lukeš, D., Skarnitzl, R. 2013. Formant Contours in Czech Vowels: Speaker-discriminating Potential. *INTERSPEECH* Lyon, 3181–3186.

[15] Morrison, G.S. 2009. Likelihood–ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *J. Acoust. Soc. Am.* 125(4), 2387–2397.

[16] Sjölander, K., Beskow, J. 2000. WaveSurfer - an open source speech tool. *Proc. 6th ICSLP* Beijing, 464−467.

[17] McDougall, K. 2006. Dynamic features of speech and the characterization of speakers: toward a new approach using formant frequencies. *Int J Speech Lang La* 13(1), 89–126.

[18] McDougall, K., Nolan, F. 2007. Discrimination of speakers using the formant dynamics of/u:/in British English. *Proc. 16th ICPhS* Saarbrücken, 1825–1828.

[19] Zuo, D., Mok, P.P.K. 2015. Formant dynamics of bilingual identical twins. *J Phonetics*, 52, 1–12.

[20] Cao, H., Wang, Y., Li, J. 2017. Effect of speaking rate on the formant dynamics of triphthongs. *J Tsinghua Univ: Nat Sci Ed* 57(9), 958–962. (in Chinese)

[21] Tabachnick, B.G., Fidell, L.S. 2013. *Using Multivariate Statistics (6th)* Boston: Pearson.

[22] Fant, G., Bavegard, M. 1997. Parametric model of VT area functions: vowels and consonants, In: *Speech, Music and Hearing Laboratory-Quarterly Progress and Status Report* Stockholm, 1–21.

[23] Takemoto, H., Adachi, S., Kitamura, T., et al. 2006. Acoustic roles of the laryngeal cavity in vocal tract resonance. *J. Acoust. Soc. Am.* 120(4), 2228–2238.

[24] Kitamura, T., Honda, K., Takemoto, H. 2005. Individual variation of the hypopharyngeal cavities and its acoustic effects. *Acoust Sci Tech* 26(1), 16–26.

[25] Dang, J., Honda, K. 1997. Acoustic characteristics of the piriform fossa in models and humans. *J. Acoust. Soc. Am.* 101(1), 456–465.