The Speaker and Language Recognition Workshop (Odyssey 2022)
28 June - 1 July 2022, Beijing, China

FORMANT DYNAMICS OF CHINESE COMPOUND VOWELS WITH IMPLICATIONS FOR FORENSIC SPEAKER IDENTIFICATION

[1,2]Jintao Kang, [1]Aijun Li, [2]Jingyang Li

[1]The Institute of Linguistics, Chinese Academy of Social Sciences
[2]Institute of Forensic Science, Ministry of Public Security, China
kangjintao@cifs.gov.cn

ABSTRACT

This study investigates the speaker-discriminatory potential of the Standard Chinese compound vowels in their formant dynamics. The first four formants were extracted from spontaneous speech materials. Two different methods (general polynomials and Legendre polynomials) of fitting formant trajectories of the 8 compound vowels from 85 male speakers were used to compute coefficients as input to a random forest classifier. Besides, original values (averages and raw data) of these formant frequencies were also input to the above classifier as benchmarks. The results show that high-frequency formants (F3 and F4) display a greater discriminatory power compared to low-frequency formants (F1 and F2) in [ai], [ia], [iou], and [uei] while low formants display greater power in the other 4 nasal compound vowels. As for curve fitting methods, Legendre coefficients perform better than the general coefficients in the random forest classifier, while the means perform worst. Furthermore, formants from [ia], [in] and [iou] are better at distinguishing speakers than other 5 compound vowels.

1. INTRODUCTION

The demand for interpretability of forensic evidence makes the formant features of vowels still one of the most frequently assessed parameters in forensic speaker identification (FSI). As mentioned by [1], vowel segments are considered to deliver information of three different dimensions: the linguistic information which conveys what was said, the social information which conveys the background of the speaker, and the idiosyncratic information. In terms of vowel formants, most studies cited by the widely-known review [2] show that research methodology concentrated on static features of vowel formants may be appropriate in monophthongs, but their conclusions are insufficient to characterize compound vowels like diphthongs and triphthongs, which have multiple articulation targets.

There are two main ways to obtain the dynamic characteristics of compound vowel formants. One method is to select measurement points based on pronunciation targets, such as two measurement points for diphthongs and three measurement points for triphthongs. The problem with this approach is that it is difficult to obtain enough information to characterize the dynamics using two or three data points. Furthermore, it is hard to decide where their targets are or where to measure for some, if not many, vowels. Another approach is to measure multiple points equidistantly or un-equidistantly along formant trajectories and then perform curve-fitting on these measurement points. The advantages of this approach are obvious. One is that a unified framework can be used without having to adapt to different types of vowels, the other is that there are more data points and richer information can be utilized. Therefore, many previous studies [3, 4, 5, 6, 7, 8, 9] have adopted the second approach and proved that dynamic features do have better discriminating power than static features in speaker identification.

The rationality of the above researches seems obvious: the phonetic realization of the formant dynamics, especially in compound vowels, would be strongly subject to the specific implementation, which is hidden in trajectories, of the acoustic targets by the speaker. As Nolan suggested in [10], "the imprint of an individual's speech mechanism (language, articulatory habits, and vocal tract anatomy combined) will be found to lie more in dynamic descriptions than in static descriptions."

As for methods of fitting formant trajectories, many researchers [3, 4, 5, 6, 7] used general polynomials with linear, quadratic, and cubic functions as basic elements for their naturalness and ease. Although the bases of general polynomials are linearly independent, they are not orthogonal, which leads to completely different results each time the order of the curve-fitting is changed. Segundo [9] also used the Discrete Cosine Transform technique to compute coefficients, but the interpretation of each coefficient is not easy.

To solve the above problems, this study investigates the dynamic acoustic properties of 8 compound vowels of Standard Chinese, [ai], [an], [ia], [iæn], [in], [iŋ], [iou], and [uei] and proposes to use Legendre polynomials, the important members of the orthogonal polynomial family, to fit formant trajectories.

2. MATERIAL AND METHODS

2.1. Speakers and speech material

The total number of speakers in this study was 85. They were all male speakers selected from RASC863 (Regional Accented Speech Corpus funded by National 863 Project) [11] and RASC863-G2, a Chinese speech corpus with 10 regional accents of Shanghai, Guangzhou, Chongqing, Xiamen, Changsha, Luoyang, Nanchang, Nanjing, Taiyuan and Wenzhou respectively. There are more than 1000 male speakers in the corpus.

As the name of the corpus suggests, every speaker in the corpus speaks Standard Chinese with a certain degree of accent due to the influence of their native dialect. In China, the accent of Standard Chinese is categorized into three levels and each level can be divided into two degrees, where A-1 is the best and C-2 is the worst. To make sure the compound vowels in this study are consistent and close to what they should be, each speaker should have an accent category of B1 or above.

10.21437/Odyssey.2022-55

In order to ensure sufficient input for subsequent classification tasks, each speaker should have at 10 tokens of each compound vowels above. We choose 85 male speakers who meet these conditions. The ages of these speakers ranged between 18 and 50 years old at the time of recording.

The content of speech material for each speaker is a 4–5-minute spontaneous recording on a specific topic, so, we have 85 recordings in this study. All speech recording files were in 16000 Hz, 16 bit, wav format.

2.2. Annotation

All raw data used in this study were extracted using Praat scripting, which must be based on annotations. Speech transcriptions of RASC863 and RASC863-G2 are accessible. However, no ready-made annotation suitable for compound vowels research is available. Montreal Forced Aligner [12] is used to generate syllable-level annotations automatically based on speech transcriptions.

We manually checked 10 annotation files randomly selected, and found no boundary limit errors worth fixing. Therefore, annotations have not been modified in any way.

Figure 1 shows an example of textgrid file generated by Montreal Forced Aligner.

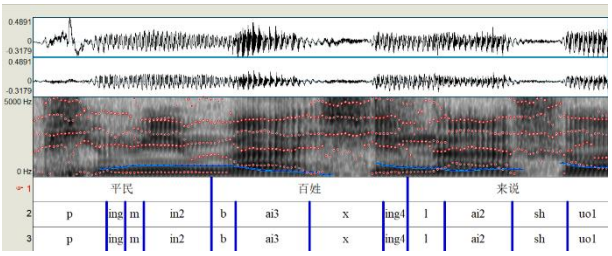

Figure 1. *Sample syllable alignment with tones generated by MFA for* ***“平民百姓来说”***

2.3. Measurements

2.3.1. Acoustic Analysis

Praat (version 6.1.42) [13] was used to perform acoustic analysis. We adopted a popular setting for extracting formant frequencies from male speakers (25 ms analysis window, 2.5 ms time step, pre-emphasis from 50 Hz, burg algorithm, maximum formant frequency of 5000 Hz, and 5 formants). For each vowel, the total duration was divided into 15 equal parts and only the first four formant frequencies were tracked and averaged at each part using a Praat script. That is to say, there are 60 data points for each vowel. Only some of the results were checked by one author using FFT spectral slices but no anomaly was corrected because the amount of data was large and the accuracy of automatic annotation was high in most cases. Table 1 and figure 2 show the formant frequencies of [ai] from a Changsha speaker.

Table 1. F1-F4 frequencies of an [ai] from speaker Changsha_male_001

F1	F2	F3	F4
726.22	1539.09	2714.68	3605.34
735.18	1518.39	2738.83	3689.48
784.16	1533.79	2742.13	3707.11
824.49	1646.06	2781.66	3711.74
811.14	1643.95	2768.68	3717.42
804.70	1562.29	2694.85	3699.64
802.93	1560.74	2566.91	3687.02
802.98	1645.96	2521.37	3685.76
805.66	1724.98	2573.67	3675.47
804.86	1673.51	2666.40	3659.67
773.62	1455.16	2533.86	3526.82
724.74	1299.00	2279.22	3435.97
707.25	1251.78	2264.78	3488.63
697.87	1163.03	2256.92	3505.14
1020.50	1888.27	2274.75	3468.10

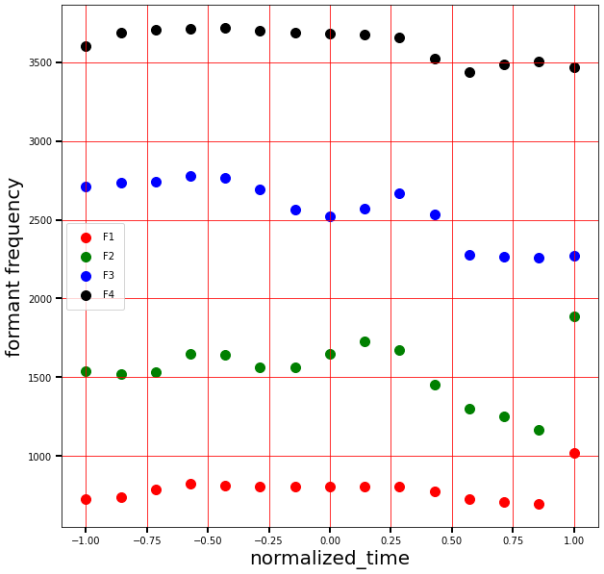

Figure 2. *F1-F4 scatterplot of that [ai] from speaker Changsha_male_001*

2.3.2. Curve Fitting

After formant frequencies were obtained, two different ways of fitting trajectories were performed: general polynomials orthogonal polynomials. These procedures are used to transform original data points into coefficients of these polynomials, thus realizing dimension reduction as well as grasping the dynamics of trajectories.

The first way of trajectory fitting approximates formant frequencies using the general polynomial function of different degrees, which has the form below,

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + a_2 x^2 + a_1 x + a_0 \quad (1)$$

where the highest order of x is the order of the polynomial function and a_n represents its coefficient. For example, a 4th order polynomial function is shown in equation (2),

$$a + bx + cx^2 + dx^3 + ex^4 \quad (2)$$

where a, b, c, d, e are its coefficients.

The second method of trajectory fitting follows the same principle as the first one, which minimizes the residual sum of squares between original values and predicted values. However, terms of the general polynomial function are not orthogonal, which makes them not perfect for many statistical analysis techniques.

Orthogonal polynomials can solve this problem. Orthogonal polynomials are classes of polynomials $\{p_n(x), n = 0,1,\ldots\}$ defined over a range [a,b] that obey an orthogonality relation,

$$(p_n, p_m) = \int_a^b \rho(x) p_n(x) p_m(x) dx = \begin{cases} 0, n \neq m \\ A_n, n = m \end{cases} \quad (3)$$

where $\rho(x)$ is the weight function.

Legendre polynomials are important members of orthogonal polynomials with the form as shown in equation (4):

$$P_n(x) = \frac{1}{2^n \cdot n!} \cdot \frac{d^n}{dx^n}[(x^2 - 1)^n], n = 0,1,2,\ldots \quad (4)$$

which is defined in the interval [-1, 1].

The first five degrees of Legendre polynomials are shown as equation (5) to equation (9).

$$P_0(x) = 1 \quad (5)$$

$$P_1(x) = x \quad (6)$$

$$P_2(x) = \frac{3}{2}x^2 - \frac{1}{2} \quad (7)$$

$$P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x \quad (8)$$

$$P_4(x) = \frac{35}{8}x^4 - \frac{15}{4}x^2 + \frac{3}{8} \quad (9)$$

Their graphs (except P_0) are depicted in Figure 3.

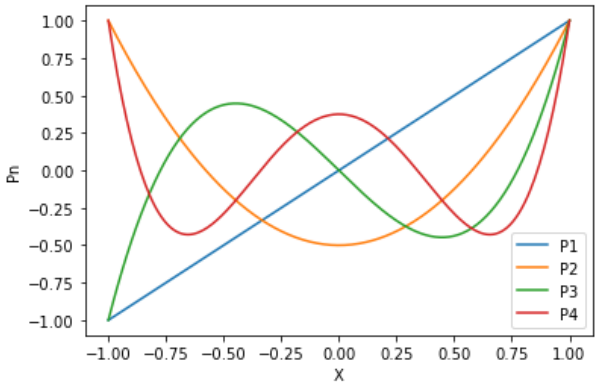

Figure 3. *Graphs of the first five Legendre polynomials generated by matplotlib*

For each formant trajectory f(x), we approximate it by an m-th order Legendre polynomial in the form of equation (10),

$$f(x) = \sum_{i=0}^{m} c_k P_k \quad (10)$$

where c_k is its k-th order coefficient, P_k is k-th order Legendre polynomial, and m is the highest order. Generally, P_0 stands for the mean of values, P_1 stands for the first derivative of the trajectory, P_2 stands for the second derivative, etc.

In this study, curve fitting and the calculation of coefficients are performed by NumPy, a package for scientific computing with Python.

Curve fitting using polynomials or Legendre polynomials is not a new thing in speech science. [4] reached a high precision with three degrees of general polynomials using linear discriminant analysis, [14] combined general polynomial coefficients with other features as input to SVM, [15, 16, 17, 18, 19] all used specifically Legendre coefficients to represent dynamics of fundamental frequencies and gained improved results in various tasks. Furthermore, as demonstrated in [20], general polynomials and Legendre polynomials will produce the same curve for the fitted data, as illustrated in Figure 4.

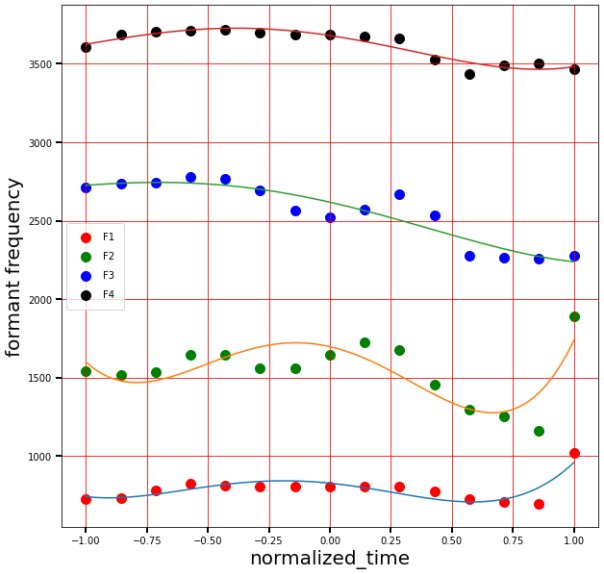

Figure 4. *Fitting curves of data points in table 1 by four degrees of general polynomials and Legendre polynomials are in full accord*

However, their coefficients are different, as shown in table 2 and table 3.

Table 2. general polynomials coefficients for data points in table 1

Coeff	F1	F2	F3	F4
a	827.15	1697.34	2618.30	3680.54
b	-157.54	-354.57	-351.94	-234.09
c	-369.16	-1232.65	-201.93	-247.16
d	265.71	422.41	108.40	162.77
e	392.44	1203.99	66.25	120.04

Table 3. Legendre polynomials coefficients for data points in table 1

Coeff	F1	F2	F3	F4
0	782.58	1527.26	2564.24	3622.16
1	1.89	-101.13	-286.90	-136.42
2	-21.85	-133.77	-96.76	-96.18
3	106.28	168.96	43.36	65.11
4	89.70	275.20	15.14	27.44

2.4. Random Forest

"Random forest" is an efficient algorithm introduced by Breiman [21]. The random forest combines a certain number of decision trees [22] in a single prediction model and is a bright gemstone in the field of ensemble learning. It can be applied to regression or classification problems. In classification problems, the algorithm outputs the final result based on the voting results of all decision trees, and thus provides greater accuracy. One of the advantages of the random forest algorithm is that it is insensitive to multi-collinearity in the input data and to variables that do not contribute to the classification strength. In this study, this is important since we will not pre-set any position on which formant will be more important in differentiating speakers, and it is assumed that formant features may be highly correlated. The "random" in random forest refers to the random sampling of variables and data entries in the learning process.

The random forest algorithm was chosen in this study for two reasons: (1) It performs well in many classification tasks in theory and practice. (2) It offers a good feature selection indicator, which is very useful in showing the relative importance of each feature in this study.

The scikit-learn toolkit (v1.0.2) [23] will be used to implement the random forest algorithm in this study.

3. RESULTS

As described in section 2.4, we used a random forest classifier implemented by the scikit-learn toolkit to perform speaker recognition using formant features. For this task, we set the number of trees in the forest to 100, the criterion as "gini" and min_samples_split to 5.

F_1 score was chosen as the metric, which is defined as equation (11),

$$F_1 = 2 * \frac{precision * recall}{precision + recall} = \frac{TP}{TP + \frac{FP + FN}{2}} \quad (11)$$

where TP, FP and FN are the number of true positives, false positives and false negatives classified by the model.

Table 4 shows the F_1 scores of the eight compound vowels from three kinds of input (averages, general coefficients, Legendre coefficients).

Table 4. F_1 scores of three kinds of input (mean values, general coefficients, Legendre coefficients) by the random forest classifier

Vowels	Means	General Coeffi	Legendre Coeffi	Original Data
[ai]	0.11	0.26	0.32	0.41
[an]	0.07	0.25	0.31	0.39
[ia]	0.1	0.38	**0.42**	**0.52**
[iæn]	0.08	0.23	0.29	0.38
[in]	0.12	0.37	0.41	**0.52**
[iŋ]	0.09	0.24	0.32	0.44
[iou]	**0.13**	0.37	0.41	**0.52**
[uei]	0.12	**0.39**	0.4	**0.51**
Average	0.1	0.31	0.36	0.46

As is shown in table 4, the Legendre coefficients perform best in the three kinds of input, which get higher f_1 scores than general coefficients in all speaker recognition tasks based on each compound vowel respectively. Mean values perform worst in these tasks with a 0.1 f_1 score, which is still better than a random picking (about 1.2% in the 85-speaker corpus). Not surprisingly, the best result came from the original data, which did not undergo any dimension reduction.

Furthermore, we can find from results that some compound vowels are better in discriminating speaker. Raw data from [ia], [in], [iou] and [uei] have reached scores higher than 0.5. Even in the form coefficients, their scores are still higher than the other four vowels.

In a random forest classifier, the feature importance can be measured as the average impurity decease computed from all decision trees in the forest.

Figure 4 shows the feature importance values and ranks from Legendre coefficients of [ai].

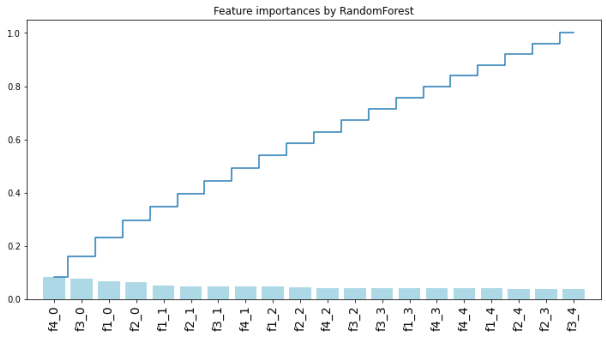

***Figure 5.** Feature importance based on mean decrease in impurity from [ai]'s Legendre coefficients*

As shown in figure 5, the first coefficients of F4 (f4_0) and F3 (f3_0) are the most important features in differentiating speakers. However, not all vowels show the same tendency. In our study, only [ai], [ia], [iou], and [uei] display a greater discriminatory power in terms of feature importance in higher formants compared to lower formants. Table 5 shows details.

Table 5. The importance ranking of each formant in discriminating speakers with random forest model

Vowels	1st	2nd	3rd	4th
[ai]	F4	F3	F1	F2
[an]	F1	F2	F3	F4
[ia]	F3	F4	F1	F2
[iæn]	F1	F4	F2	F3
[in]	F2	F1	F4	F3
[iŋ]	F1	F2	F3	F4
[iou]	F4	F1	F2	F3
[uei]	F4	F3	F2	F1

As a result, nasal vowels are different from other vowels in their importance rankings of formants' discriminatory power.

4. CONCLUSION AND DISCUSSION

In this study, we compared two curve fitting methods for capturing formant dynamics of eight Chinese compound vowels. As figure 4 shows, their visual effects in fitting

trajectories are the same. However, the results in table 4 demonstrate that coefficients from these two kinds of methods perform differently in the random forest model used in this study.

A possible reason is that the general polynomials and the Legendre polynomials have different bases, which makes coefficients from the latter less correlated, or more independent and identically distributed. Machine learning models, such as random forest, love iid. Another possible reason, which might not be mentioned by others, is Legendre coefficients fit the model used in the study better, for reasons unknown.

As for the original data's best performance, the reason may be that, for a sufficiently complex model, like the random forest model in this study, the more informative the data, the better the results. The model can find extra information, which is lost in the process of dimension reduction.

The differences in formants' importance rankings of discriminatory power between nasal vowels and non-nasal vowels may originate from inherent flaws in formant measurement using LPC. The LPC is an all-pole model, but the nasal vowels bring zero points from the nasal cavity, making the formant frequencies generated by LPC unreliable.

The implications of the present study's results on FSI practice may come in two ways. Firstly, there is an everlasting demand regarding the identification of robust parameters in this time-pressing and time-consuming field. If some temporal intervals, such as [ai], [ia] and [iou], are more powerful in discriminating speakers than others, we should invest more time in these segments. Secondly, formant features are usually considered to have higher interpretability in court for many studies in various disciplines and their close relation to physiological characteristics. A deeper understanding of formants' potential and limits will be helpful in identifying explanatory factors accounting for the speaker's idiosyncratic features. Therefore, the results of this study provide valuable references for other researches and will promote the development of practice.

5. ACKNOWLEDGEMENT

This study was partially sponsored by the China Institute of Forensic Science, under contract number 2019JB032.

6. REFERENCES

1. P Ladefoged, "Information Conveyed by Vowels," THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, vol. 29, no. 1, 1957, [Online]. Available: http://acousticalsociety.org/content/terms.

2. R. D. Kent and H. K. Vorperian, "Static measurements of vowel formant frequencies and bandwidths: A review," Journal of Communication Disorders, vol. 74. Elsevier Inc., pp. 74–97, Jul. 01, 2018. doi: 10.1016/j.jcomdis.2018.05.004.

3. M. Jessen, "Speaker-specific information in voice quality parameters," International Journal of Speech Language and the Law, vol. 4, no. 1, pp. 84–103, May 2013, doi: 10.1558/ijsll.v4i1.84.

4. J. Li, L. Wang, J. Cui and X. Wang, "Formant dynamics of Standard Chinese diphthongs," in Proceedings of PCC2012, pp. 161–164, May. 2012.

5. D. Loakes, "A forensic phonetic investigation into the speech patterns of identical and non-identical twins," International Journal of Speech Language and the Law, vol. 15, no. 1, Jul. 2008, doi: 10.1558/ijsll.v15i1.97.

6. K. McDougall, "Speaker-specific formant dynamics: An experiment on Australian English /aI/," International Journal of Speech, Language and the Law - Forensic Linguistics, vol. 11, no. 1, pp. 103–130, Jun. 2004, doi: 10.1558/sll.2004.11.1.103.

7. K. McDougall, "Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies," International Journal of Speech Language and the Law, vol. 13, no. 1, pp. 89–126, Feb. 2007, doi: 10.1558/ijsll.v13i1.89.

8. D. Zuo and P. P. K. Mok, "Formant dynamics of bilingual identical twins," Journal of Phonetics, vol. 52, pp. 1–12, Sep. 2015, doi: 10.1016/j.wocn.2015.03.003.

9. E. San Segundo and J. Yang, "Formant dynamics of Spanish vocalic sequences in related speakers: A forensic-voice-comparison investigation," Journal of Phonetics, vol. 75, pp. 1–26, Jul. 2019, doi: 10.1016/j.wocn.2019.04.001.

10. F. Nolan, "The 'telephone effect' on formants: a response," International Journal of Speech Language and the Law, vol. 9, no. 1, pp. 74–82, Mar. 2007, doi: 10.1558/ijsll.v9i1.74.

11. A. Li, Z. Yin, T. Wang, Q. Fang, and F. Hu, "RASC863 - A Chinese Speech Corpus with Four Regional Accents," 2004.

12. M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in Interspeech 2017, Aug. 2017, pp. 498–502. doi: 10.21437/Interspeech.2017-1386.

13. P. Boersma and D. J. M. Weenink, "PRAAT, a system for doing phonetics by computer," GLOT International, vol. 5, no. 9, pp. 341–347, Dec. 2001, [Online]. Available: https://www.researchgate.net/publication/208032992

14. J. Hou, Y. Liu, T. F. Zheng, J. Olsen, and J. Tian, "Multi-layered Features with SVM for Chinese Accent Identification," in Proceedings of ICALIP2010, Nov. 2010, pp. 25–30.

15. L. Tan and M. Karnjanadecha, "PITCH DETECTION ALGORITHM: AUTOCORRELATION METHOD AND AMDF," 2003.

16. L Ruiter, "How useful are polynomials for analyzing intonation?" in Proceedings of Interspeech2008, 2008, pp. 785–788.

17. D. J. Weenink, "IMPROVED FORMANT FREQUENCY MEASUREMENTS OF SHORT SEGMENTS," 2015.

18. X. Mao, B. Zhang, and Y. I. Luo, "SPEECH EMOTION RECOGNITION BASED ON A

HYBRID OF HMM/ANN," in Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications, Aug. 2007, pp. 367–370.

19. C.-Y. Lin and H.-C. Wang, "LANGUAGE IDENTIFICATION USING PITCH CONTOUR INFORMATION," in Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, pp. 601–604.

20. Jincai CHANG, Long ZHAO, and Qianli YANG, "Data Fitting Method Based on Orthogonal Polynomials," Journal of Hebei Polytechnic University, vol. 33, no. 4, pp. 79–84, 2011.

21. L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5–32, 2001.

22. S. B. Kotsiantis, "Decision trees: a recent overview," Artificial Intelligence Review, vol. 39, no. 4, pp. 261–283, Apr. 2013, doi: 10.1007/s10462-011-9272-4.

23. F. Pedregosa FABIANPEDREGOSA et al., "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," 2011. [Online]. Available: http://scikit-learn.sourceforge.net.