Proceedings of Meetings on Acoustics

Volume 50 http://acousticalsociety.org/

183rd Meeting of the Acoustical Society of America

Nashville, Tennessee

5-9 December 2022

Psychological and Physiological Acoustics: Paper 2pPP6

Inferring pitch from coarse spectral features

Danni Ma, Neville Ryant and Mark Liberman

University of Pennsylvania, Philadelphia, PA, USA; dannima@seas.upenn.edu; nryant@gmail.com; markyliberman@gmail.com

Fundamental frequency (F0) has long been treated as the physical definition of "pitch" in phonetic analysis. But there have been many demonstrations that F0 is at best an approximation to pitch, both in production and in perception: pitch is not F0, and F0 is not pitch. Changes in the pitch involve many articulatory and acoustic covariates; pitch perception often deviates from what F0 analysis predicts; and in fact, quasi-periodic signals from a single voice source are often incompletely characterized by an attempt to define a single time-varying F0. In this paper, we find strong support for the existence of covariates for pitch in aspects of relatively coarse spectra, in which an overtone series is not available. Thus linear regression can predict the pitch of simple vocalizations, produced by an articulatory synthesizer or by human, from single frames of such coarse spectra. Across speakers, and in more complex vocalizations, our experiments indicate that the covariates are not quite so simple, though apparently still available for more sophisticated modeling. On this basis, we propose that the field needs a better way of thinking about speech pitch, just as celestial mechanics requires us to go beyond Newton's point mass approximations to heavenly bodies.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0001696

1. INTRODUCTION

Pitch, as defined by the American National Standard on Acoustical Terminology (ANSI/ASA S1.1-2013),[33] is the attribute of auditory sensation by which sounds are ordered on the scale from low to high. Pitch describes how human ears and brains interpret acoustic signals. Since it is a perceptual attribute, in order to characterize pitch in a quantitative fashion, researchers have been using fundamental frequency (F0) as the physical surrogate of pitch, where F0 refers to the greatest common divisor of all the involved frequencies of harmonics.

It is widely recognized that the key predictor of the pitch of a sound is its ***periodicity***. The definition of F0 has an assumption that overtones are perfectly periodic in the frequency domain, and likewise acoustic signals are perfectly periodic in the time domain. However, this is not necessarily true in practice. For example, all string instruments produce overtones that are slightly sharper than ideal harmonics, due to effective shortening of the string at higher frequencies.[19,36] The minor inharmonicity makes the sound waveforms not perfectly periodic. But since the deviation is only a few percent, the waveforms are called *quasi-periodic*.

So is human voice. From an articulatory point of view, F0 is the frequency at which vocal folds vibrate in voiced sounds. Voiced speech signals consist of periodic signals and random perturbations. The random component is mostly jitter and shimmer, which are quantified as the cycle-to-cycle variations of F0 and amplitude.[32] Human voice is also deemed quasi-periodic with small perturbations.

Given the nature of musical sound and human voice, the notion of F0 does not entirely apply to sound in the real world. Researchers in the past have investigated this phenomenon; e.g., Shepard[29] discusses *Shepard tone illusion*, suggesting that perceived pitch can not be adequately represented by a purely rectilinear scale because of the circularity of pitch space.[3,5,6] A substantial number of literature has discussed the impact of voice quality on facilitating pitch perception. Swerts and Veldhuis[30] and Honorof and Whalen[8] is early work that presented the potential dependency of F0 on voice quality. Lee,[16] Allen and Oxenham,[2] Kuang and Liberman[12,13,15] and Kuang et al.[11] further discover that changes in spectral shape affect people's judgements of relative pitch height. Besides spectral slope, Kuang and Liberman[14] suggests that vocal fry, as another voice quality cue, would bias listeners to perceive a lower pitch. More generally, phonation is an important correlate of tone perception for various tonal languages.[4,7,37,38] There are abundant other non-f0 cues for pitch perception, including sound intensity[21] and temporal envelope,[9] to name a few.

On the other hand, pitch does not exclusively depend on F0 in production. Zhang[40] proposes that vocal folds tension plays an important role in the control of pitch in voice production. Similarly, Titze[35] analyzes the correlation between different types of phonation and pitch production. Later, Roubeau et al.[26] and Kuang[10] show that voice quality is closely associated with pitch production in both speaking and singing voices. Specifically, lowest pitch production is associated with creaky voice, while the highest pitch production is associated with falsetto or whistle. Rhee et al.[25] also proves that voice quality helps the acquisition and production of Mandarin tones.

It is unsurprising that changes in pitch, as we will point out in this paper later, have covariates other than changes in the spacing of overtones, or the spacing of repetition in the time domain. Prior work has provided sufficient evidence, either implicitly or explicitly, of this argument. Ryant et al.[27,28] show that highly accurate Mandarin tone classification is possible using only MFCC features, when both F0 and overtone series are absent. Lin et al.[17] and Zhang et al.[39] discuss the task of Mandarin tone classification and pitch range estimation respectively, but both work found that a model achieves better results with the addition of spectral features, compared to the one that uses prosodic features only.

In order to better validate and understand the existence of latent information related to pitch in spectral features, we design experiments that track pitch on acoustic data where only F0 is varied. We study both synthetic and real human voice data. The goal is to see whether predicting F0 estimates from spectral features only is possible, and how difficult it would be. We believe that the results will suggest a promising

direction for better modeling prosodic features in human voice.

2. HUMAN STIMULI

A. DATA COLLECTION

For the human experiments, we recorded two human subjects (one male and one female). They were instructed to sinusoidally vary pitch for three vowels: /ɑ/, /æ/ and /i/. For each vowel, subjects repeated this process multiple times until at least 5 minutes of usable utterances had been produced. Subjects were encouraged to vary the period, pitch range, and starting frequency of their oscillations from utterance to utterance, but to be careful to not vary these parameters within an utterance. Subjects were recruited from the student population at the University of Pennsylvania (age range: 20-30 years) and neither had professional voice training.

Table 1: Statistics of pitch from human stimuli.

Subject	Property	max	min	avg	std
Female	**`Period`** (s)	1.17	0.40	0.5838	0.1605
	`Pitch` (Hz)	747.07	67.87	300.80	113.37
	`Duration` (s)	20.39	5.15	8.24	2.47
	`Pitch range` (Hz)	192.53	20.07	49.90	39.71
Male	**`Period`** (s)	1.32	0.40	0.7223	0.2375
	`Pitch` (Hz)	363.02	64.37	149.38	43.39
	`Duration` (s)	20.59	3.63	12.04	4.23
	`Pitch range` (Hz)	85.90	10.01	24.66	17.29

Table 1 shows some major statistics of the collected stimuli, where **`Pitch`** is the F0 estimate acquired from RAPT algorithm[31] as implemented in the Speech Signal Processing Toolkit (SPTK).[1] The following parameters are used:

- **`wind_dur`** = 10 ms
- **`min_f0`** = 60 Hz
- **`max_f0`** = 800 Hz for female, 400 Hz for male

In the following experiments, we treat F0 estimate from RAPT as gold pitch. **`Period`** is the period of sinusoidal oscillation produced by subjects. For example, the period of sinusoid shown in Fig. 1 is approximately 0.8 seconds. **`Duration`** is the duration of one utterance, and **`Pitch range`** is the difference between the highest and lowest pitch in each period. Figure 2 gives a more detailed distribution of pitch values and period. As expected, the pitch range of the female speaker is more widely spread out; while the male speaker generally produces slower-varying vocalizations.

B. ANALYSIS

Figure 1 plots the F0 estimates, spectrogram, and 40-D Mel-frequency cepstral coefficients (MFCCs) (Section 3.1 for details) corresponding to one of the female /ɑ/ productions. The spectrogram is plotted using a 20 ms long Hanning window. Dynamic range is set to 45 dB, and pre-emphasis is 0 dB/Octave. For

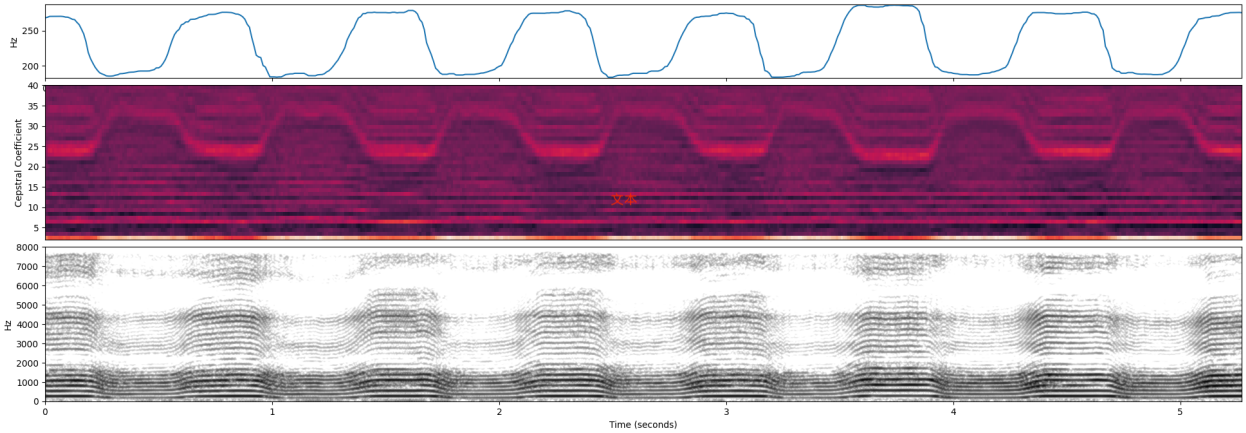


Figure 1: An example production of /ɑ/ from the female speaker. **Top*****: F0 estimate from RAPT.*** **Middle*****: Plot of cepstral coefficients (Section 3.1 for details).*** **Bottom*****: Log-power spectrum.***

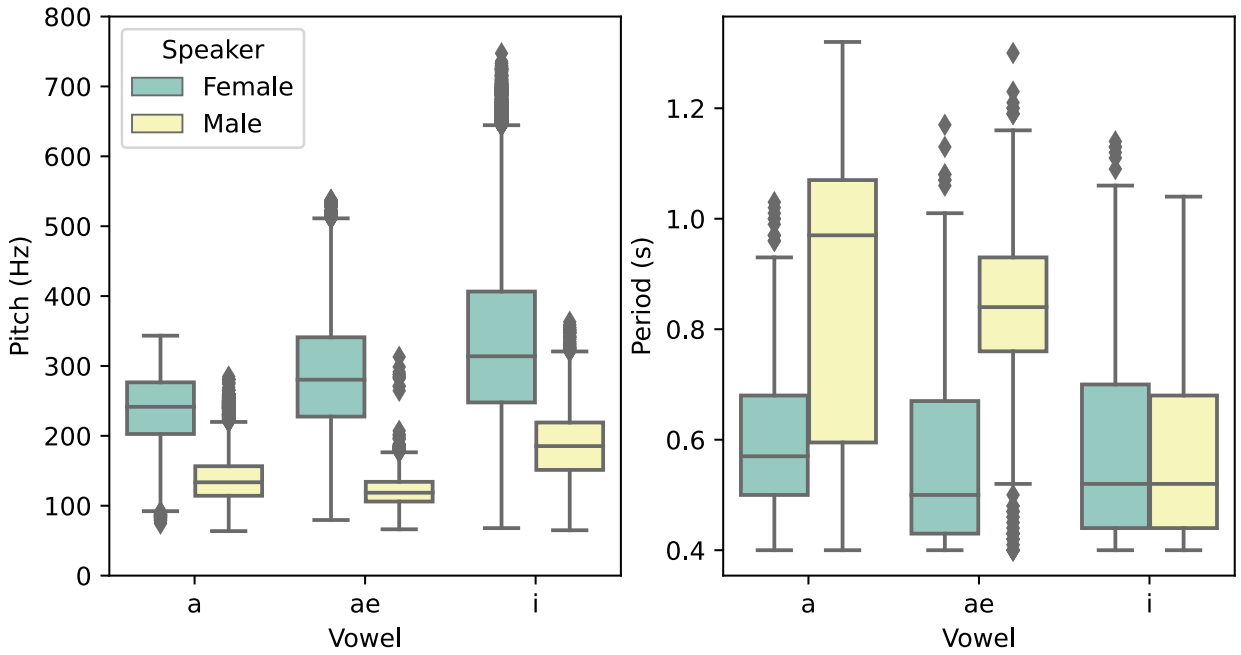


Figure 2: Distribution of pitch range (left) and period (right) for each combination of speakers and vowels.

this production, the oscillation in F0 is clearly reflected in both the spectrogram and the MFCCs, particularly in the higher coefficients. Similar patterns are revealed in productions of the other vowels and for the male speaker.

Given the interesting patterns that emerged from visual inspection of MFCCs for isolated utterances, we compute the Pearson's correlation[23] between F0 and each dimension of cepstral coefficients, plotted in Fig. 3. While all correlations are statistically significant, the correlations are particularly strong for 10th-28th cepstral coefficients. This pattern is even stronger for the female speaker.

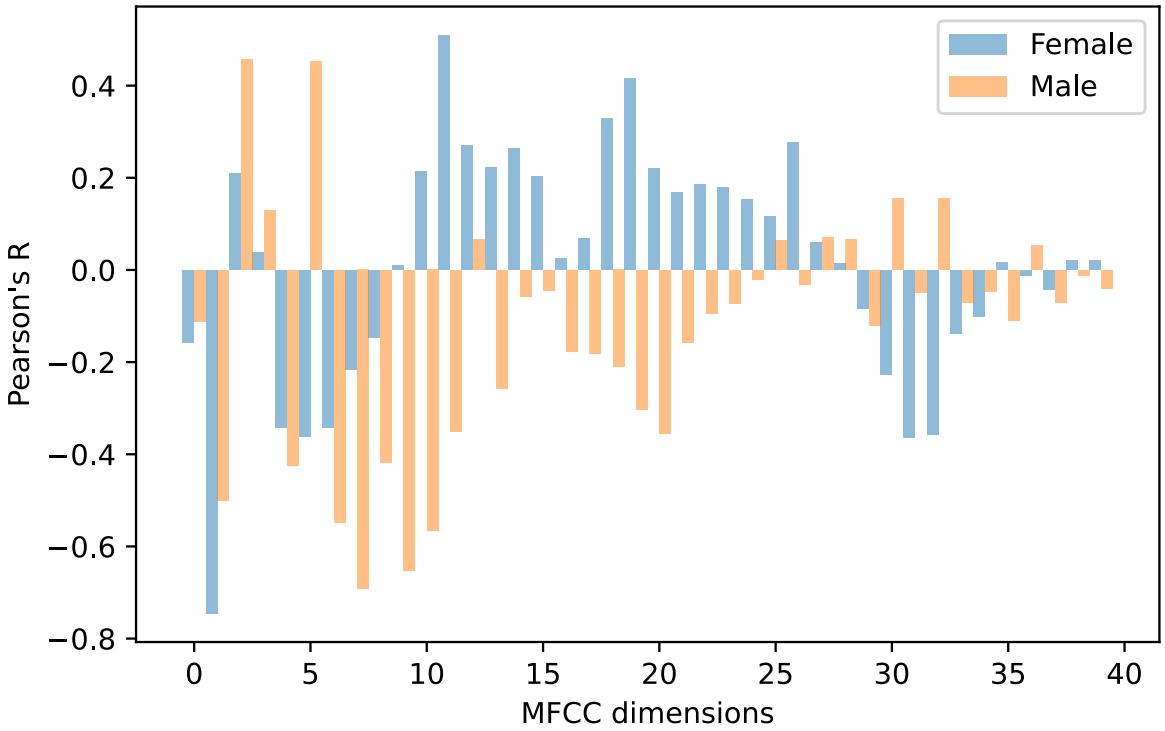


Figure 3: Pearson correlation between F0 and each dimension of cepstral coefficients. The correlation is obtained from all the vowel stimuli collected from each speaker.

3. EXPERIMENTS

A. METHOD

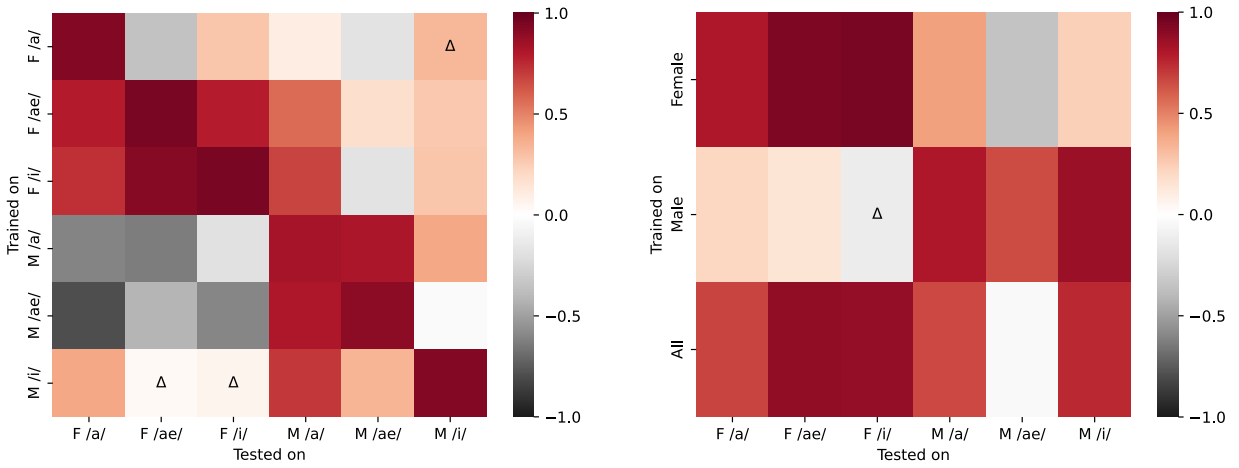
Task – Each type of stimuli are partitioned into training and test set with an 80/20 split. The task is to predict time functions of F0. In order to alleviate the affect of different pitch ranges from different speakers (as shown in Fig. 2), F0 values are transformed from Hz to semitones using the 5th percentile of each speaker as the base.[22] Semitones that are transformed from gold pitch are referred to as gold semitones in the results. We do a pairwise combination of training and test sets, whose results constitute a 6×6 matrix shown in Fig. 4(a). The number that each cell represents is the average of 10 runs, and in each run the training/test split is different. This is to reduce the impact of randomness, as we have a fairly small dataset.

Acoustic features – The collected recordings are resampled to 16kHz, and stored as mono 16-bit audio. Then we use 40-D MFCCs as the features of the input acoustic signals. MFCC coefficients are extracted using *librosa*[18] with a 10 ms step size and a 35 ms analysis window. No overtone series is explicitly present in this representation.

Model – We utilize linear regression without using an L2 loss as implemented by the `LinearRegression` class in *sklearn*[24] for all regression tasks.

Evaluation – Root-mean-square error (RMSE) and Pearson's R are used as evaluation metrics to measure the difference and correlation respectively between the ground truth and predicted semitone values.[1]

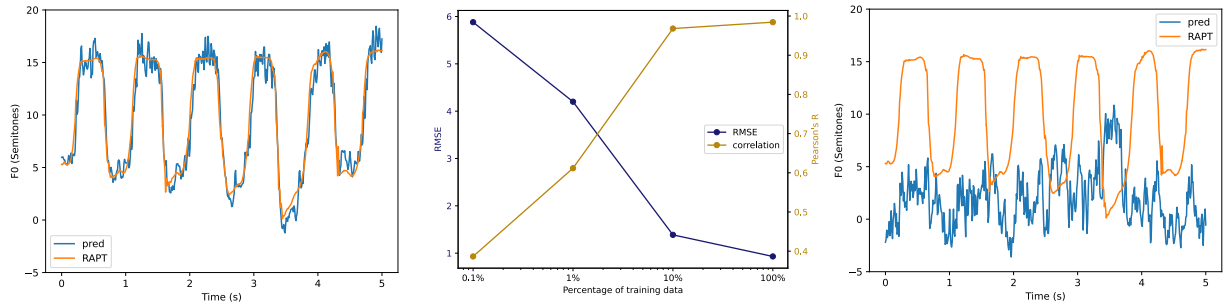
[1]The code and data are publicly available at https://github.com/dannima/InferringPitch



(a) Train on one vowel one speaker

(b) Train on all the vowel stimuli

Figure 4: Pearson correlation coefficient between the predicted and gold semitone values for the experiments involving human data. Each cell represents one train/test set combination with the y-axis defining the training set and the x-axis the test set. Darker/lighter colors indicate stronger/weaker correlation. (a) Training on a single vowel from a single speaker. (b) Training on* all *vowels from either a single or both speakers.* Abbreviations*: F: female speaker, M: male speaker, △: result in cell is* not *statistically significant ($p > 0.05$)



(a) Train and test on the same speaker.

(b) Different amounts of training data.

(c) Train and test on different speakers.

Figure 5: Pitch prediction of an excerpt of the vowel /æ/ stimuli from the female speaker. (a) Predicted (pred) vs gold (RAPT) semitones of a sample utterance. (b) The RMSE and Pearson correlation between the gold pitch and predicted pitch based on the amount of training data. Full training data is* 4 minutes*. (c) Same as (a), except that the linear regression model is trained on the same vowel stimuli from the male speaker, and tested on the female speaker.

B. RESULTS

Figure 5 shows the most salient results of the experiments. When training and testing on the same vowel stimuli from the same speaker, even with such a small amount of training data and simplest regression model, MFCC spectral parameters are able to predict the pitch from real human voice. The predictions in Fig. 5(a) roughly fit the gold pitch tracks, except being a little jagged in peaks and troughs.

More interestingly, if the training data is reduced to 10%, RMSE remains almost the same, meaning that around ***25 seconds*** of MFCC features is enough for the model to predict the pitch well. More variations of training data and its corresponding performance can be found in Fig. 5(b). Knowing that RMSE is greatly affected by the pitch range of training data, we also plot Pearson's r between the gold pitch and

the prediction. The conclusion still holds. This is strong evidence of the argument that pitch has spectral covariates other than the overtone series.

But Fig. 5(c) indicates that the covariates are not that simple. The linear regression model struggles to recover F0 if tested on different kinds of materials. Additional experiments show that by adding contexts to MFCC features and change to a more complicated model, such as multi-layer perceptron regressor, will improve the performance in both inter-speaker and cross-speaker settings. It is plausible that compared to linear models, MLP can better extract prosodic information from MFCC features.

4. SYNTHETIC STIMULI

We replicate the same task on synthetic data and it yields similar results. The conclusion is even stronger – It is possible to recover much more complicated oscillations of pitch using the combination of single-frame MFCCs and a linear regression model. Below are the details of experiments and findings.

A. DATA GENERATION

We generate synthetic data using Pink Trombone[342], an articulatory synthesizer by varying the F0 of the glottal wave while keeping all other control parameters fixed. Five-second duration utterances are generated by sampling the non-pitch control parameters, and the F0 contour is determined according to one of two mechanisms:

Sinusoidal stimuli. F0 of the glottal wave is varied according to a sinusoid with the following form:

$$\text{F0} = 172 * \sin(\alpha * t + \phi) + 232 \tag{1}$$

where $\alpha \in [1.7227, 8.6133]$ and $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. The resulting F0 values lie in the range [60 Hz, 404 Hz]. Both α and ϕ are randomly selected within respective ranges. Therefore, each of the stimuli has different pitch range, period and phase shift to increase the diversity of inputs.

Complicated stimuli. F0 is generated as the superposition of two sinusoidal oscillations A_t and ω_t:

$$A_t = \sin(\alpha_1 * t + \beta_1) \tag{2}$$

$$\omega_t = \cos(\alpha_2 * t + \beta_2) \tag{3}$$

$$\text{F0} = 172 * A_t * \sin(\omega_t * t + \phi) + 232 \tag{4}$$

where $\alpha_1, \alpha_2 \in [0.8613, 3.4453]$ and $\beta_1, \beta_2, \phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

In order to simulate the settings in human voice, we generate 60 five-second long audio files for each kind of stimuli, resulting in five minutes worth of synthetic data for each mechanism. Then they are divided into training/test set with an 80/20 split as well. All the other experimental settings, including features, model, and task, are the same as described in Section 3.1.

B. RESULTS

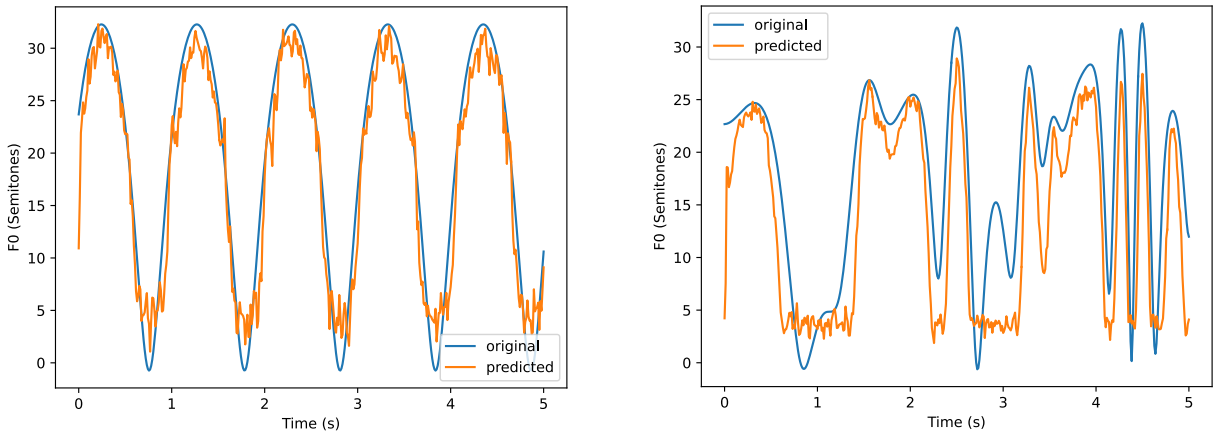
Two samples of the regression results are shown in Fig. 6. Results in Fig. 6(b) are surprisingly good. A simple linear regression model is able to mostly recover the intricate shape of the input.

In order to have a quantitative assessment of the performance, Table 2 lists the RMSE of the training and test set for the synthetic pitch prediction task. Much to our surprise, errors of different kinds of stimuli are close. It is reasonable to hypothesize that, regardless of the complexity of input signals, spectral features contain enough information to predict pitch.

[2]Specifically, using the following Python re-implementation: https://github.com/dkadish/pynktrombone

Table 2: RMSE of synthetic pitch prediction.

Stimuli	Training	Test
`Sinusoidal`	3.079	2.733
`Complicated`	2.592	2.364



(a) Sinusoidal stimuli of pitch

(b) Complicated stimuli of pitch

Figure 6: Synthetic stimuli from Pink Trombone and the predicted pitch.

5. DISCUSSION

A. MEASURE OF PITCH

In this study, we use F0 as the ground truth of pitch, despite pointing out that F0 and pitch are not exactly the same. We argue that the claim should be orthogonal to how pitch is represented in the experiments. F0 is used because it is closely correlated with pitch, and it is also one of the cheapest measures of pitch to obtain. This work shows that F0 can be (mostly) recovered solely based on MFCCs, without the presence of overtone series. The results are sufficient enough to corroborate the covariation between spectral features and pitch.

B. COMPLICATED COVARIATES

The success of single-frame MFCC features predicting pitch is unexpected. Experiments presented so far provide solid evidence of the existence of covariates. But many aspects of the covariates remain obscure. We are particularly interested in the reason for performance degradation in Fig. 5(c).

We explore more diverse experimental settings in Fig. 4 to seek an answer. Training and testing on vowels produced by the same speaker obviously work better than testing on vowels produced by a different speaker. But it is also observed that training on both speakers can mitigate disadvantages in cross-speaker testing.

Additionally, pairs of vowels that are located closer in IPA vowel chart, i.e., have similar articulatory features, yield better results when predicting pitch. For example, within the same speaker, models trained on /ɑ/ have better performance when testing on /æ/ than on /i/.

The failure in Fig. 5(c) could be due to the lack of training data, but a more interesting hypothesis is not ruled out – Covariates are not simple in human speech. There might be higher-order interactions

among properties of spectra regarding speakers, recording conditions, phonetic contexts etc. We defer more investigation of this hypothesis for future research.

C. BETTER REPRESENTATIONS OF PITCH

The existence of covariates also encourages researchers to find something other than F0 to describe pitch. An ideal representation of pitch would address current concerns of pitch tracking. It would not suffer from frequent doubling/halving errors.[20] And it should be homomorphic to human perception. Existing pitch tracking methods have the problem that a minuscule change in the input can cause substantial change in the output. The lack of continuity is generally not favorable for the interpretability of a representation.

Currently, the spiral properties of pitch and human auditory illusions introduce all the complication of representing pitch gracefully. In the future, we will work more in the direction of better modeling prosodic features.

6. CONCLUSION

This paper shows a surprising phenomenon that coarse spectral representations are able to infer pitch without the presence of overtone series, by using even the simplest form of regression models. Therefore, the results present a strong support for the existence of covariates for pitch in coarse spectra. We validate the argument by experimenting on both human voice and synthetic articulations. And we further discover that, the covariates might be in a complex form and are affected by speakers and phonetic contexts. Altogether, this study provides a further understanding of prosodic features, and suggests a future research direction of better characterizing speech pitch.

REFERENCES

[1] Speech Processing Toolkit (SPTK). https://github.com/sp-nitech/SPTK.

[2] Emily J Allen and Andrew J Oxenham. Symmetric interactions and interference between pitch and timbre. *The Journal of the Acoustical Society of America*, 135(3):1371–1379, 2014.

[3] Ira Braus. Retracing one's steps: An overview of pitch circularity and Shepard tones in european music, 1550–1990. *Music Perception*, 12(3):323–351, 1995.

[4] Marc Brunelle. Tone perception in northern and southern Vietnamese. *Journal of Phonetics*, 37(1):79–96, 2009.

[5] Edward M Burns. Circularity in relative pitch judgments for inharmonic complex tones: The Shepard demonstration revisited, again. *Perception & Psychophysics*, 30(5):467–472, 1981.

[6] Diana Deutsch. The paradox of pitch circularity. *Acoustics Today*, 7:8–15, 2010.

[7] Marc Garellek, Patricia Keating, Christina M Esposito, and Jody Kreiman. Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America*, 133(2):1078–1089, 2013.

[8] Douglas N Honorof and Douglas H Whalen. Perception of pitch location within a speaker's F0 range. *The Journal of the Acoustical Society of America*, 117(4):2193–2200, 2005.

[9] Ying-Yee Kong and Fan-Gang Zeng. Temporal and spectral cues in Mandarin tone recognition. *The Journal of the Acoustical Society of America*, 120(5):2830–2840, 2006.

[10] Jianjing Kuang. Covariation between voice quality and pitch: Revisiting the case of mandarin creaky voice. *The Journal of the Acoustical Society of America*, 142(3):1693–1706, 2017.

[11] Jianjing Kuang, Yixuan Guo, and Mark Liberman. Voice quality as a pitch-range indicator. In *Proceeding of Speech Prosody*, pages 1061–1065, 2016.

[12] Jianjing Kuang and Mark Liberman. The effect of spectral slope on pitch perception. In *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[13] Jianjing Kuang and Mark Liberman. Influence of spectral cues on the perception of pitch height. In *Proceedings of The International Congress of Phonetic Sciences (ICPhS)*, 2015.

[14] Jianjing Kuang and Mark Liberman. The effect of vocal fry on pitch perception. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5260–5264. IEEE, 2016.

[15] Jianjing Kuang and Mark Liberman. Integrating voice quality cues in the pitch perception of speech and non-speech utterances. *Frontiers in Psychology*, 9:2147, 2018.

[16] Chao-Yang Lee. Identifying isolated, multispeaker mandarin tones from brief acoustic input: A perceptual and acoustic study. *The Journal of the Acoustical Society of America*, 125(2):1125–1137, 2009.

[17] Ju Lin, Wei Li, Yingming Gao, Yanlu Xie, Nancy F Chen, Sabato Marco Siniscalchi, Jinsong Zhang, and Chin-Hui Lee. Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features using extended recognition networks. *Journal of Signal Processing Systems*, 90(7):1077–1087, 2018.

[18] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference (SciPy)*, volume 8, pages 18–25, 2015.

[19] James Anderson Moorer. *On the segmentation and analysis of continuous musical sound by digital computer*. Stanford University, 1975.

[20] Kathleen Murray. A study of automatic pitch tracker doubling/halving "errors". In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001.

[21] John G Neuhoff and Michael K McBeath. The Doppler illusion: the influence of dynamic intensity change on perceived pitch. *Journal of Experimental Psychology: Human Perception and Performance*, 22(4):970, 1996.

[22] Julia Parish-Morris, Mark Liberman, Neville Ryant, Christopher Cieri, Leila Bateman, Emily Ferguson, and Robert T Schultz. Exploring autism spectrum disorders using HLT. In *Proceedings of the Conference Association for Computational Linguistics Meeting*, volume 2016, page 74. NIH Public Access, 2016.

[23] Karl Pearson. Notes on regression and inheritance in the case of two parents. In *Proceedings of the Royal Society of London*, volume 58, pages 240–242, 1895.

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[25] Nari Rhee, Aoju Chen, and Jianjing Kuang. Integration of spectral cues in the development of mandarin tone production. In *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 3135–3138. Australasian Speech Science and Technology Association Inc., 2019.

[26] Bernard Roubeau, Nathalie Henrich, and Michèle Castellengo. Laryngeal vibratory mechanisms: the notion of vocal register revisited. *Journal of voice*, 23(4):425–438, 2009.

[27] Neville Ryant, Malcolm Slaney, Mark Liberman, Elizabeth Shriberg, and Jiahong Yuan. Highly accurate Mandarin tone classification in the absence of pitch information. In *Proceedings of Speech Prosody*, volume 7, pages 673–677, 2014.

[28] Neville Ryant, Jiahong Yuan, and Mark Liberman. Mandarin tone classification without pitch tracking. In *Proceedings of The International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4868–4872. IEEE, 2014.

[29] Roger N Shepard. Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America*, 36(12):2346–2353, 1964.

[30] Marc Swerts and Raymond Veldhuis. The effect of speech melody on voice quality. *Speech Communication*, 33(4):297–303, 2001.

[31] David Talkin and W Bastiaan Kleijn. A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, 495:518, 1995.

[32] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis – jitter, shimmer and HNR parameters. *Procedia Technology*, 9:1112–1122, 2013.

[33] Acoustical Terminology. Standard ansi/asa s1. 1-2013. *American National Standard Institute*, 2013.

[34] Neil Thapen. Pink Trombone. https://dood.al/pinktrombone, 2017.

[35] Ingo R Titze. A framework for the study of vocal registers. *Journal of Voice*, 2(3):183–194, 1988.

[36] Caroline Traube and Julius O Smith. Estimating the plucking point on a guitar string. In *Proceedings of the Conference on Digital Audio Effects (DAFx'00), Verona, Italy*, pages 153–158, 2000.

[37] Ruo-Xiao Yang. The phonation factor in the categorical perception of Mandarin tones. In *Proceedings of The International Congress of Phonetic Sciences (ICPhS)*, pages 2204–2207, 2011.

[38] Kristine M Yu and Hiu Wai Lam. The role of creaky voice in Cantonese tonal perception. *The Journal of the Acoustical Society of America*, 136(3):1320–1333, 2014.

[39] Qi Zhang, Chong Cao, Tiantian Li, Yanlu Xie, and Jinsong Zhang. Pitch range estimation with multi features and mtl-dnn model. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pages 939–943. IEEE, 2018.

[40] Zhaoyan Zhang. Mechanics of human voice production and control. *The Journal of the Acoustical Society of America*, 140(4):2614–2635, 2016.