Investigating prosody in music and speech

Yundu Wang[1], *Elinor Payne*[2]
[1]Guildhall School of Music & Drama, London, UK
[2]Phonetics Laboratory, University of Oxford, and St Hilda's College, Oxford, UK
yundu.wang@stu.gsmd.ac.uk, elinor.payne@phon.ox.ac.uk

Abstract

We investigated the speech and musical performances of six classical pianists, of native Mandarin Chinese and English language backgrounds, comparing the prosodic properties in their speech with temporal expressivity in their piano performances. We expected intra-language consistency. Results, while mixed, suggest both intra-language and intra-speaker consistency, which implies that individual, expressive (performative) ability affects both speech and music.

Index Terms: musical performance, prosody, English, Mandarin Chinese, rhythm, production, speech

1. Introduction

Studies on the cognitive parallels between speech and music have primarily focused on perception. Research suggests that the two domains share 'basic processing mechanisms', including the ability to absorb and learn sound categories, to perceive regularities from rhythmic and melodic sequences, to integrate elements such as words and musical tones into syntactic structures, and to extract meaning from and emotional responses through sound (cf. [1]).

Cross-domain research on performance is limited, but studies suggest that classical musicians exhibit expressive nuances in their performances that are similar to prosodic elements in speech. Examples include phrase-final lengthening to mark boundaries and changes in intensity and duration to mark prominence ([2][3]).

Combinative studies that are both cross-domain and cross-linguistic are rare, but promising. A speech rhythm index (nPVI) ([4]), based on capturing degree of variability in vocalic intervals in speech was used to examine the musical themes of composers from language backgrounds claimed to belong to differing linguistic rhythm groups ([5][6]). It was determined that the nPVI scores for music corresponded with associated language scores; music for which a higher nPVI was obtained (e.g. American, English, and Swedish) were written by composers whose native language was also associated with a high linguistic nPVI, whereas French, Italian, and Spanish music had significantly lower nPVI. While the neat typological rhythmic classification of languages, has met with some criticism, not least because rhythm metric scores have been shown to vary as a function of speech task and individual performance as much as between languages, there are broad distributional differences between languages that suggest linguistic properties do play some role in shaping, or constraining the rhythmic properties of speech. Here we investigate whether linguistic background may also influence musical expressivity.While some empirical work on cross-lingusitic influences upon musical performance exists (e.g. [7][8][9]), none have involved both ecological validity (use of repertoire from the Western classical canon in a performance setting) and analyses of the participants' speech.

It has been proposed that certain linguistic properties, such as syllable-structure, quantitative vowel reduction, and stress-induced lengthening, contribute to differing rhythmic impressions in speech, giving rise to a distinction between so-called 'stress-timing' and 'syllable-timing' ([10]) - terms which have persisted even though it is now understood that the basis of these perceptual differences is not isochrony of units such as the syllable of prosodic foot. Research suggests that read Mandarin Chinese speech has a lower nPVI_V and a higher %V than read English speech, and that native Mandarin speakers of English have a lower Varco_V and higher %V than native English speakers ([11]). While the tone system of Mandarin and stress system of English are fundamentally different, it has been claimed that Mandarin has some form of prominence alternation = with 'neutral', toneless syllables having reduced duration and schwa-like qualities [12][13] . In terms of boundaries, research suggests that English phrase-final lengthening occurs only on the syllable that actually abuts the boundary, while lengthening in Mandarin is reported to begin earlier (i.e. during the pre-final syllable) and the boundary syllable itself is shorter than in English (in spontaneous speech; [15]).

This study investigates whether a performer's native language influences or informs the 'musical prosody' ([3]) of his or her performance. Specifically, it seeks to answer the question, 'Is language influence reflected through the expressive execution of rhythm and phrasing during performance?' This preliminary investigation has the following hypothesis: Within musical performance, we predict differences in the degree of timing variability (above and beyond what is prescribed by the composed notation) between language groups.

2. Methodology

2.1. Participants

The participants were six classical pianists, residing, at the time of the experiment, in London, UK. Three participants were Mandarin Chinese (*Putonghua)* native speakers with L2 English. The remaining three were monolingual English speakers, including two American and one British. All three Mandarin speakers were from mainland China and had lived in the UK for a period of 1-9 years. All participants completed a language proficiency questionnaire to build fluency profiles, which included time spent in an English-speaking country as

10.21437/SpeechProsody.2018-111

well as self-ratings of language fluency in English and Mandarin.

2.2. Materials and elicitation

The English speech data, read by all six participants, consisted of productions of an excerpt from 'A Serious Case' ([16]). The Mandarin speech data, read only by the three Mandarin speakers, consisted of read productions of an excerpt from 'Outside the Window' (窗外) ([17]). The musical data consisted of performances of an excerpt from 'Rosemary', a work for piano solo composed by the English composer Frank Bridge (1879-1941). Each participant was asked to play the excerpt in two styles: i) mechanically (without expression) and ii) with expression. Studies have shown that such directions alter the amount of a performer's expressive nuances ([18][19]). Both versions were recorded for each participant. Auditory impressions of the expressive performances of the six participants were made by both the author (a professional classical pianist) and co-author, prior to analyses. A professional classical pianist was recruited to conduct a blind listening test of the music recordings.

Both speech and music recordings were made in the recording studio of the Guildhall School of Music & Drama in London. The speech recordings were made in a vocal booth, equipped with a Neumann U87 microphone. The pianists performed on a Steinway Model B grand piano, recorded with a pair of DPA 4011 (cardioid) microphones. All recordings were made using ProTools software.

2.3. Labelling

Consonant and vocalic intervals were segmented from the waveform and spectrogram and start-points and end-points labelled on a syllabic tier in *Praat.* Vocalic and consonantal segmentation were carried out with reference to standard criteria: placement of boundaries between vocalic and consonantal intervals was guided primarily by the presence of a sudden, significant drop in amplitude and a break in the formant structure, particularly F2. Marking of the consonant onsets was facilitated by various cues, according to the manner of the consonant. In addition to syllable boundaries, syllables were also labelled according to segmental content and according to level of prominence. For prosodic prominence, syllables were labelled as belonging to one of three levels: i) unstressed; ii) stressed; iii) nuclear stressed. Both intonational and intermediate phrase boundaries were identified and marked. Syllables, as well as vocalic and consonantal intervals, were categorized according to phrase position (initial, medial, or final).

The music recordings were visualised through spectrograms and note onsets (time instances) were labelled manually using *Sonic Visualiser* ([20]), a free software system designed for the analysis of musical sound files. As with the speech analysis, labelling was aided by spectrographic and audio information. Timings of onsets were labelled on three separate beat levels (time instance layers): i) semi-quaver beat; ii) quaver beat; iii) crotchet beat. Bar timings were labelled on a separate layer. Time instances of the melody in the right hand (notes in the treble clef of Figure 1) were also labelled. The differences between the time points of successive time instances yielded the interonset intervals (IOIs). IOIs for each beat level were extracted from both the mechanical and expressive versions and labelled according to bar and beat numbers. IOIs for the melody layer were labelled according to note durations (*DQ* for dotted quaver, *SQ* for semiquaver, *C* for crotchet, and *M* for minim).

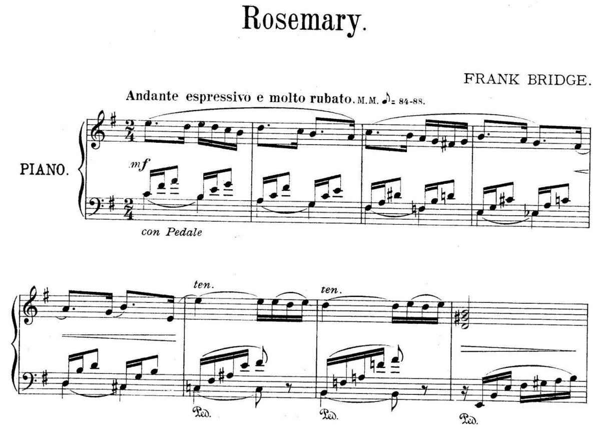

Figure 1: *The piano score of 'Rosemary' from 3 Sketches, H. 68 (1906)*

2.4. Analysis

Durations of syllabic, consonantal, and vocalic intervals were extracted using a *Praat* script. The following rhythmic measures were calculated for four English sentences by each speaker, and mean values calculated for each speaker: %V, ΔV, VarcoV, and nPVI_V (this paper focuses on English speech - only the syllabic intervals of the Mandarin Chinese sentences were extracted for speech rate analysis). For the purposes of this study (see [5] and [6]), only vocalic durations were analysed, since these are comparable with note durations. Music durations of each performance were analysed and the mean, standard deviation, nPVI_V, and VarcoV were calculated. For this preliminary study, only durations of the quaver beat level were targeted (the quaver pulse allows for detailed durational analysis without resorting to note-by-note durational extraction).

In both speech and music analyses, duration values for phrase-final units were excluded from the calculations due to their possible distortive effect on the measures. In speech, pre-final and final phrase syllables were excluded. This is to control for any pre-final syllable lengthening than may occur in the Mandarin speakers. In music, penultimate and ultimate bars was excluded for the same purpose. While degree of final lengthening is of interest in itself, it can be analysed separately from measures of more holistic timing variability.

Speech-rate and overall tempo were calculated for speech (both English and Mandarin Chinese sentences) and music, respectively. Speech-rate (syllables per second or sps) was calculated manually by dividing the total number of syllables by the total length of the sentence. Again, the final and pre-final syllables were excluded to control for phrase-final lengthening. Large pauses (often triggered in speech by the presence of commas and semicolons, and demarcating phrase boundaries) were excluded from the total length. Interestingly, all three Mandarin speakers of English had a number of smaller pauses between syllables, and it was decided that these should be included in the total length sum, as such pauses are interpreted as a characteristic of the three Mandarin speakers (when speaking L2 English). The scores of four sentences were averaged for each speaker. In music, the average quaver-note duration was calculated by dividing the total duration by the total number of quaver beats (16 quaver beats were analysed for each performance). It was necessary to exclude beats that were involved in phrase-final lengthening and prominence cues (conspicuous slowing down, lengthening and/or delaying of notes to highlight importance) (see [21] for reference). From the average quaver-note duration, the average metronome speed was determined (quavers per minute or qpm).

Table 1: *Speech rate (sps) and Metronome speed*

Participant	English	Chinese	qpm
Mandarin 1	4.17	5.05	96
Mandarin 2	4.61	4.33	77
Mandarin 3	4.87	5.49	100
English 1	6.34	-	117
English 2	6.50	-	85
English 3	5.91	-	80

2.5. Results

2.5.1. Speech rate and overall tempo

Analysis of speech rate of read English sentences showed a language-based grouping of results, with native Mandarin speakers having a slower speech rate than native English speakers. This is a common phenomenon of second-language speakers ([11]), although interestingly the Mandarin sentences were also read with slower speech rates.

Analysis of overall tempo showed weak language-based grouping: the Mandarin group scored an overall average of 91 qpm while the English group scored 94 qpm. Between language groups, English 2 and 3 had slower tempi than Mandarin 1 and 2.

A detailed cross-comparison of speech rate with overall tempo did not show a consistent speaker-based grouping: English 1 had the fastest tempo, while English 2 had the fastest speech rate (n.b., English 3 had the slowest speech rate and overall tempo). Interestingly, some consistency remained within the Mandarin group: Mandarin 3 had the fastest speech rate and overall tempo. Mandarin 2 had the slowest overall tempo as well as speech rate, but only in Mandarin speech. In English speech, Mandarin 1 had the slowest speech rate, which may be contributed by a low self-rated fluency of English speech (more in the Discussion).

2.5.2. nPVI_V and VarcoV

Table 2: *nPVI_V and VarcoV of speech, with overall averages.*

Participant	nPVI_V	VarcoV
Mandarin 1	61.41	50.58
Mandarin 2	43.35	42.84
Mandarin 3	35.09	37.05
M. average	46.62	43.49
English 1	44.63	34.04
English 2	53.41	47.6
English 3	48.15	44.48
E. average	48.73	42.04

Table 3: *nPVI_V and VarcoV of music.*

Participant	nPVI_V	VarcoV
Mandarin 1	10.43	13.25
Mandarin 2	11.76	12.57
Mandarin 3	8.24	7.96
English 1	14.21	22.07
English 2	11.51	14.02
English 3	11.22	10.39

Analysis of nPVI_V and VarcoV in English speech showed weak language-based grouping. Mandarin 1 had the highest nPVI_V of both language groups, which suggests that there may be significant variation across speakers within any one language.

Analysis of nPVI and Varco scores for IOIs showed moderate consistency of language-based grouping. The average nPVI score was 10.14 in the Mandarin group, compared to 12.31 in the English group. The average Varco score was 11.26 in the Mandarin group, compared to 15.49 in the English group. Only one participant (English 3) had a score lower than those of the Mandarin speakers (further discussion below).

Auditory impressions of the expressive performances of the six participants were made by both the author (a professional classical pianist) and co-author, prior to analyses. A further set of impressions were made by a professional classical pianist in a blind listening session. A score ranging from 1 to 5 was given to each participant, based on the amount of durational variation heard in the recording, with 1 being the least and 5 being the highest. The Varco scores of individuals represented well the impressions of both the author and co-author, as well as the scores of the third listener.

Table 4: *Scores of blind test and VarcoV*

Participant	Blind	VarcoV
Mandarin 1	3	13.25
Mandarin 2	2	12.57
Mandarin 3	2	7.96
English 1	5	22.07
English 2	4	14.02
English 3	2	10.39

2.5.2. *Spectrograms*

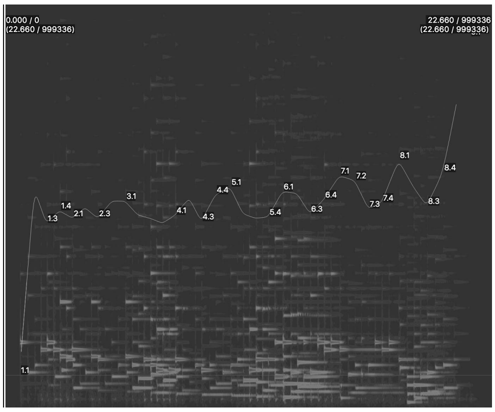

Figure 2: *Mandarin 3 expressive version*

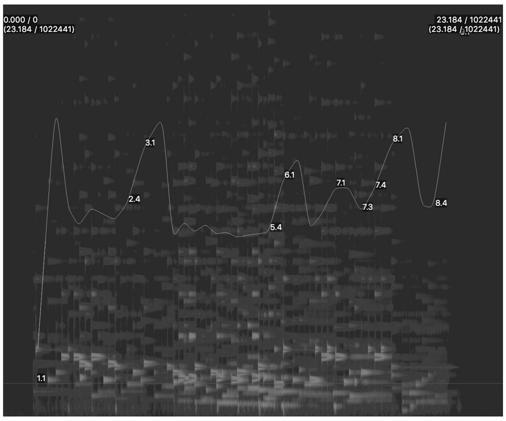

Figure 3: *English 1 expressive version*

Spectrograms made in *Sonic Visualiser* allowed for visual comparisons of the durational variability between performances. Line graphs with a curved plot type show variation between successive quaver-beat durations. Curves rise upwards as durations lengthen and fall as they shorten. Comparisons between the performances of Mandarin 3 (lowest Varco score) and English 1 (highest Varco score) show significant contrast. Mechanical versions of each participant show the extent of durational contrast between mechanical and expressive performance. The mechanical versions of Mandarin 3 and English 1 are quite similar; this is true for all six performances.

3. Discussion

This preliminary investigation focused on two properties of production in both speech and music: timing variability and rate. Results of a comparison between speech rate and tempo were inconclusive for either language-based grouping or intra-speaker consistency. Overall tempo scores were higher for two Mandarin speakers than for two English speakers. Although English 1 had the fastest overall tempo in performance, this was not reflected in speech rate. English 2 had the fastest speech rate, but not a significantly fast tempo. It is interesting to note that English 1 was British, while English 2 and English 3 were from eastern and western United States, respectively. These results bring to attention the extent of prosodic variability between dialects of a language, including regional dialects.

The nPVI_V and VarcoV measures in speech revealed inconclusive language-based grouping. Mandarin 1 had the highest scores of both measures out of all the participants. English 1 had the second lowest nPVI_V and the lowest VarcoV score out of all the participants. This is inconsistent with the results of both the durational and impressionistic analyses of English 1's musical performance. It has been suggested in studies of nonnative stress and rhythm production that Mandarin speakers of English would lengthen stressed syllables more, rather than shorten unstressed syllables (as would be expected for native English speakers). Also, discrepancies between listeners' impression of nonnative speech and acoustic measures may be a result of slower speech rate and 'selective lengthening' (see [11]). As for the results of English 1, subsequent analysis of consonantal durations, stress and boundaries will be conducted.

Results of nPVI and Varco measures in music are more suggestive of the possibility of language-based grouping. The overall average of nPVI music scores for Mandarin speakers were lower than that of English speakers. Varco scores were even more supportive of language-based grouping within performances. Additionally, intra-speaker consistency was suggested by both the nPVI and Varco scores. Scores were largely consistent with listeners' impression and spectrographic comparison. The nPVI and Varco scores of Mandarin 2 were the only discrepancy. Mandarin 2 had the highest nPVI score and the median Varco score. It is possible that the tempo or speed of performance affected the nPVI score, since Mandarin 2 had the slowest tempo (and the measure does not control for speed rate). This phenomenon could also explain the highest nPVI_V score (and slowest speech rate) for Mandarin 1's English speech.

4. Conclusions

This study has conducted preliminary acoustic measurement on the speech and musical performance of classical pianists with differing native language backgrounds, with the purpose of investigating the diversity of performative aspects in both and music performance, and possible correlations with degree of expressivity. Results suggest that while speech rate and overall tempo are not consistent with language-based or speaker-based grouping, they may affect respective nPVI_V measures for both speech and music. This study also suggests, based on nPVI and Varco measures in music, the possibility of both intra-language and intra-speaker consistency in speech and musical performance.

Next steps will include analysis of further data gathered on different kinds of speech (neutral sentence reading, spontaneous speech), as well as different degrees of expressivity in performance (mechanical and expressive), and analysis of intensity and duration variation between stressed and unstressed syllables.

5. Acknowledgements

We would like to thank the Guildhall School of Music & Drama for providing support and resources, as well as the participants for letting us analyse their voices and performances.

6. References

[1] Patel, A.D. (2008) *Music, Language, and the Brain.*New York: Oxford University Press.

[2] Palmer, C., Jungers, M., and Jusczyk, P. W. (2001) Episodic memory for musical prosody. *Journal of Memory and Language*. 45. pp.526-545.

[3] Palmer, C. and Hutchins, S. (2006) What is musical prosody? In Ross, B.H. (ed.). *Psychology of Learning and Motivation.* 46 (1). Amsterdam, The Netherlands: Elsevier Press. pp.245-278.

[4] Grabe, E. & Low, E. L. (2002) Durational variability in speech and the rhythm class hypothesis. In Gussenhoven, C. & Warner, N. (Eds.). *Laboratory phonology* 7 Berlin: Mouton de Gruyter. pp.515–546.

[5] Patel, A.D. and Daniele, J.R. (2003). An empirical comparison of rhythm in language and music. *Cognition.* 87. pp. B35-B45.

[6] Huron, D., and Ollen, J. (2003). Agogic contrast in French and English themes: Further support for Patel and Daniele (2003). *Music Perception*.21.pp.267–271.

[7] Ohgushi, K. (2002) Comparison of Dotted Rhythm Expression between Japanese and Western Pianist., *7th International Conference on Music Perception and Cognition,* Sydney.

[8] Sadakata, M., Ohgushi, K., and Desain, P. (2004) A cross-cultural comparison study of the production of simple rhythmic patterns. *Psychology of Music.* 32. pp.389–403.

[9] Slobodian, L.N. (2008) Perception and production of linguistic and musical rhythm by Korean and English middle school students. *Empirical Musical Review.* 3 (4).

[10] Dauer, R. M. (1983) Stress timing and syllable-timing reanalyzed. *Journal of Phonetics.* 11 pp. 51-62.

[11] Mok, P., Dellwo, V. (2008). Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. in the Speech Prosody 2008, s.n., Campinas, Brazil, pp. 423–426.

[11] Chao, Y. R. (1968). A grammar of spoken Chinese. Berkeley and Los Angeles: University of California Press.

[12] Chen, Y., Xu, Y. (2006). Production of Weak Elements in Speech – Evidence from F_0 Patterns of Neutral Tone in Standard Chinese. *Phonetic.* 63 pp.47–75.

[13] Zhang, Y. H., Nissen, S. L., and Francis, A. L. (2008) Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *The Journal of the Acoustical Society of America.* 123. pp.4498-4513.

[14] Fon, J., 2002. A Cross-Linguistic Study on Syntactic and Discourse Boundary Cues in Spontaneous Speech. Columbus, OH: The Ohio State University dissertation.

[15] Rose, C., n.d. A Serious Case. [Online] *Learn English | British Council.* Available from: https://learnenglish.britishcouncil.org/en/stories-poems/serious-case [accessed: 14 June 2017].

[16] Learn Mandarin Chinese. (2015) Elementary Level Chinese Readings: Story 窗外(Outside of the Window). [Online] Available at: http://tcfl.tingroom.com/2015/01/6421.html. [Accessed: 14 June 2017].

[18] Gabrielsson, A. (1987) Action and Perception in Rhythm and Music. *Papers Given at a Symposium in the Third International Conference on Event Perception and Action.* Royal Swedish Academy of Music.

[19] Seashore, C.E., 1936. New Vantage Grounds in the Psychology of Music. *Science.* 84.pp.517–522.

[20] Cannam, C., Landone, C., and Sandler, M. (2010) Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files. *Proceedings of the ACM Multimedia 2010 International Conference.*

[21] Repp, B.H., 1988. Patterns of expressive timing in performances of a Beethoven Minuet by nineteen famous pianists and one computer. *The Journal of the Acoustical Society of America,* 83 S120–S120.