INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Musical Speech: a New Methodology for Transcribing Speech Prosody

Alexsandro R. Meireles[1], *Antônio R. M. Simões*[2], *Antonio Celso Ribeiro*[1], and *Beatriz Raposo de Medeiros*[3]

[1]Federal University of Espirito Santo, Brazil
[2]University of Kansas, USA
[3]University of São Paulo, Brazil

meirelesalex@gmail.com, asimoes@ku.edu, antoniocelsoribeiro@gmail.com, biarm@usp.br

Abstract

Musical Speech is a new methodology for transcribing speech prosody using musical notation. The methodology presented in this paper is an updated version of our work [12]. Our work is situated in a historical context with a brief survey of the literature of speech melodies, in which we highlight the pioneering works of John Steele, Leoš Janáček, Engelbert Humperdinck, and Arnold Schoenberg, followed by a linguistic view of musical notation in the analysis of speech. Finally, we present the current state-of-the-art of our innovative methodology that uses a quarter-tone scale for transcribing speech, and shows some initial results of the application of this methodology to prosodic transcription.

Index Terms: speech prosody, speech melodies, musical notation, quarter tones

1. Introduction

It is known among linguists that speech is composed of musical elements such as speech rhythm, intonation, tonicity, and speech dynamics. Speech Prosody is the area of Linguistics that investigates this musicality. In recent years there has been an ongoing interest in this field, and since 2002 a special conference, Speech Prosody (http://isle.illinois.edu/sprosig/sp2002/), has been organized to investigate all interdisciplinary aspects of speech prosody. Although musical aspects are present in speech, researchers have rarely tried to represent speech using traditional musical notation. On the other hand, composers such as Bach, Beethoven, Schoenberg, Reich, and Janaček have represented speech using traditional scores. Indeed, Bach was one of the pioneers in musically scoring speech with the *recitativo secco*, a technique in which a singer is allowed to adopt the rhythms of ordinary speech [1, 17].

In the preface of *Pierrot Lunaire* (1912) (http://www.schoenberg.at/index.php/en/onlineshop/product/212-pierrot-lunaire-companion-paperback) Arnold Schoenberg instructed the performers to not sing the musical notes on the score. The singers, therefore, should interpret them as a representation of speech, i.e., as a "speech-melody". The rhythm should be "as if you were singing", but the speaking tone should vary the pitch continuously with falling or rising inflexions. Figure 1 shows one excerpt from this piece. It is clear from Schoenberg's own words in this preface that he intended to represent speech with musical notation but not to be an exact representation of it, for "in no way should one strive for realistic, natural speech".

Rapoport [2, 3] investigated the origins of Schoenberg's notation. According to these studies, the musical rhythm was based on the syllabic structure of the German text written by Otto Erich Hartleben, and the speech melody (*Sprechmelodie*) was based on German's intonation patterns. To prove this hypothesis, the voices of two German speakers (1 male born in 1920, and 1 young female) were recorded. These vocal analyses of speech intonation contours in spoken, read aloud by two native speakers of German, brings experimental evidence to elucidate the origin of Schoenberg's "Sprechmelodies" in German intonation.

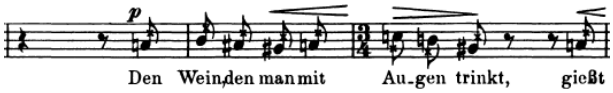

Figure 1: *Excerpt of Schoenberg's speech-melody (represented by the crosses "x" on the notes) as seen in Pierrot Lunaire op. 21. The cross represents the speech notes. Freely available at http://imslp.org/wiki/Pierrot_Lunaire,_Op.21_(Schoenberg,_Arnold).*

Rapoport [2, 3] investigated the origins of Schoenberg's notation. According to these studies, the musical rhythm was based on the syllabic structure of the German text written by Otto Erich Hartleben, and the speech melody (*Sprechmelodie*) was based on German's intonation patterns. To prove this hypothesis, the voices of two German speakers were recorded. These vocal analyses of speech intonation contours, taken from the recording of texts read aloud by two native speakers of German, bring experimental evidence to elucidate the origin of Schoenberg's "Sprechmelodies" in German intonation.

Schoenberg used speech-melody extensively in his compositions, and developed this technique, but he was not the first to use it. Engelbert Humperdinck's *Königskinder* (1897) is known as the first time *Sprechmelodie* was used. Later, Alban Berg, Schoenberg's student, used this technique in his famous operas *Wozzeck* (1922) and *Lulu* (1935).

In the realm of auditory perception, Deutsch's research [4, 5] sheds some light on how speech can result in music. The professor and her colleagues [4] have discovered that when a sentence is repeated over and over, it is perceptually interpreted as music rather than speech. In another experiment, Deutsch and colleagues [5] have found neurological evidence for this perceptual illusion. According to the latter study, "a network of brain regions associated with the perception and production of pitch sequences showed greater response when subjects listened to the song stimuli" [5, p. 5].

This musical illusion in speech may be the basis on which many composers support their songs. There are many pop songs that use standard tuning to represent speech[1], such as Frank Zappa's *Jazz Discharge Party Hats* (1983), *So happy* by

[1] These songs can be easily found on youtube.

http://dx.doi.org/10.21437/Interspeech.2017-316

Steve Vai (1984) and *Yankee Rose* by David Lee Roth and Steve Vai (1986). Vai is well-known as a virtuoso transcriber as seen in *So happy*, a song in which the long spoken text is mimicked by his guitar, similar to what happens in Zappa's song. In *Yankee Rose*, we have an example of how the guitarist talks to the vocalist using the guitar mechanics. Figure 2 shows how the guitar emulates the phrase "She's beautiful" with slides and the use of the whammy bar. We remind, as stated by Deutsch et al. [4, p. 2245], that "speech consists of frequency glides that are often steep, and of rapid amplitude and frequency transitions". Steve Vai seems to be aware of this speech characteristic since he uses these dynamics to simulate speech.

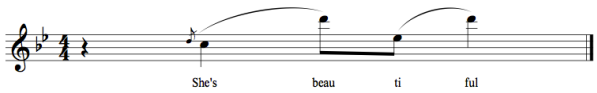

Figure 2: *Steve Vai's guitar emulating the phrase "She's beautiful".*

In the light of what has been exposed so far, it may be said that speech and music - although with different and separate functions - appear to "meet again" in different aspects of their production/perception. This is the case of composers who saw in the musical score notation a tool also feasible for speech, or the case of researchers who observed how the anchorage of different pitches in speech can be perceived by listeners as music.

2. Systematic transcription of speech with musical notation

2.1. Janáček's Theory of Speech Melodies

According to Fiehler [6], Leoš Janáček (1854-1928) developed his Theory of Speech Melodies at the time he was writing his opera Jenůfa (1903). In this same text, the author states that the composer intended to faithfully represent speech. For this reason, as happens with speech, "his melodic fragment are unstructured, phrased unconventionally, unconstrained by key or meter" [6, p. 42].

Tyrrel [7, p.793] points out that Janáček related in 1916 that "tune is created by the word, the whole melody depends thus upon the sentence, it couldn't be otherwise". Nevertheless, even though the composer always denied he fits words to an existing tune, Tyrrel [7] shows in his paper that he frequently did it when he rewrote his musical pieces.

Vainiomäki [8] shows that Wundt's psychology influenced Janáček to study speech melodies with experimental data. That's why he started to record data with Hipp's chronoscope, "with which he would measure speaking time when he was checking the time data on the melodies of speech" (p. 172). Figure 3 shows an example of Janáček's application of his speech melody to a song [8, p. 200].

Although Janáček's theory is not very instructive on how to transcribe speech melodies, we acknowledge that he tried to build a very scientific method to do it. To reinforce this perspective, we mentioned that he "included the exact concrete circumstances of the speech melody" [8, p. 170], and notated the speech melody with all information necessary for its recording such as rhythm, tempo, dynamics, and agogics [8, p. 171].

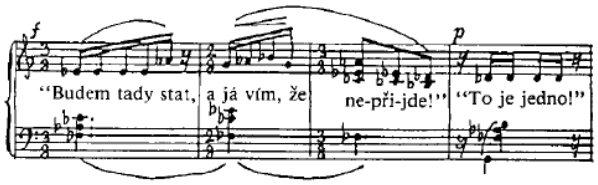

Figure 3: *Janáček's example of a music based on a speech melody*

Even though Janáček had a very good musical ear, his transcription of speech melodies was subjected to his perceptual viewpoint [8, p.244]. This drawback is recognized by the composer, who said that "speech melodies become distorted and trivialized when they are captured in notes" and that "not everyone is able to recognize or understand a speech melody in its notation" [8, p.171]. That's why Janáček's theory was not widespread as it was, for example, the dodecaphonic German school of Schoenberg.

2.2. Joshua Steele's methods on melody and measure of speech (1775)

The idea of transcribing speech using music notation comes even before Janáček's time. In 1775 Joshua Steele wrote the book "An essay towards establishing the melody and measure of speech to be expressed and perpetuated by peculiar symbols" [9]. Even though he did not have the modern instrumental tools to help him with transcription, he developed a very interesting way of notating speech with music notation. Contrary to Janáček, though, Steele's was more interested in the linguistic side of speech and his main objective was to show that "the musical expression of speech may be described and communicated in writing" [9, p. 15].

This study has one great point in common with our method, since he commented that speech should be better represented using a quarter-tone scale. This scale consists of 24 tones as follows: C, C (1/4 up), C#, C# (1/4), D, D (1/4), D#, D# (1/4), E, E (1/4), F, F(1/4), F#, F# (1/4), G, G(1/4), G#, G# (1/4), A, A(1/4), A#, A# (1/4), B, B(1/4). Figure 4 shows Steele's quarter-tone scale. In his case he chose to use 'x' to increase the frequency ¼ higher than the previous notes. So, for example, G(1/4) has 1 'x'; G#, 2 superimposed 'x'; and G# (1/4), 3 superimposed 'x'. Later we are going to see that modern quarter-tone scales use different notations.

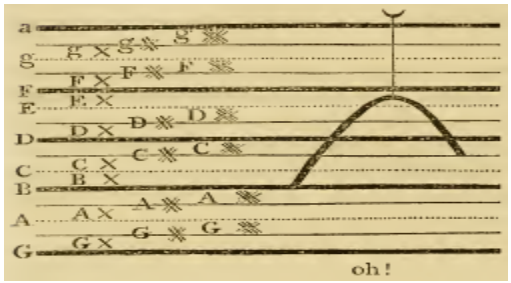

Figure 4: *Steele's representation of speech with a quarter-tone scale.*

Steele's transcription is very detailed and looks like a mix of f0 curves, music notation, and other representations of prosodic notation. Segmental symbols like notes and rests, as well as symbols representing musical dynamics were elements of Steele's detailed transcription, but to describe his notation system is beyond the scope of this paper.

3. A linguistic view on the musical aspects of speech

This paper intends to present a new methodology for transcribing speech prosody, which is inserted in a research program that proposes to develop a new model for studying speech prosody with music notation. As far as we know, no researcher has tried to fully develop a model for studying speech prosody using musical notation. Nevertheless, as seen before, some studies have transcribed speech using this method. A linguist that looked for musical elements in speech was the Russian phonologist Roman Jakobson.

Karbusický [10] consider that Jakobson's investigations of speech intonations is a line of musical ethnography in which language and music intersect as semiotic systems. In this text the author comments that Jakobson was familiar with Janáček's study of speech melodies. Also, as commented by Vainimomäki [8, p. 163], Jakobson's study "On Czech Verse, Especially in Relation to Russian Verse" investigated "natural and aesthetically functioning word material which pursues, in addition to stress and quantity, also the musical level of the affective speech act".

On section 2, we have seen the difficulties of Janáček and other composers to faithfully represent the speech melodies, which in most cases were based on subjective means. Schoenberg is also known to have problems for explaining his ideas on the exact performance of his speech scores. Boulez [11] comments "whether it is actually possible to speak according to a notation devised for singing. This was the real problem at the root of all the controversies. Schoenberg's own remarks on the subject are not in fact clear". Due to these problems, after Pierrot Lunaire, Schoenberg did not use a traditional clef in *Ode to Napoleon Bonaparte* (1942) and in *Survivor from Warsaw* (1947) he completely eliminated a specific pitch in the score, but retained the relative slides and articulations.

After presenting a short history of the transcription of speech melodies, the next sections will describe how we developed a scientific way of transcribing speech melodies, speech rhythm, and speech dynamics using a quarter-tone scale[2], i.e., a subdivision of a musical scale with 24 tones instead of the traditional 12 tones.

4. Methods

In Simões and Meireles [12] we started to develop a method for transcribing speech prosody with music notation. The transcriptions of musical notes in hertz of Brazilian Portuguese, Mexican Spanish and American English were automatically done using the software *Ableton Live 9 Suite* and then annotated in *Finale 2011*. After that, we compared the musical scores of some printed musical scores and realized that the musical transcriptions via wave to midi conversion were not very much precise, i.e., the automatic transcriptions did not match the published scores. That is why we devised a more sophisticated method of transcription based on the acoustic signal.

Due to software limitations, we could only transcribe speech using a traditional 12-tone scale. Nevertheless, by measuring the frequencies in speech, we noticed that very often the syllable's frequencies did not correspond to the frequencies in this scale. That is why we thought in expanding this scale to include smaller intervals than the semitone. At first we thought using a 48-tone scale, but as divisions smaller than the quarter tone are very difficult for listeners (even trained musicians) to perceive, we chose to use a quarter-tone scale, which is used for example in Arabic music and is also known by electric guitarists. Therefore, our scale consisted of 24 tones equally spaced (see figure 5).

In order to divide the scale in 24 tones we used the formula $f = f0 * \sqrt[24]{2}^{i}$, where f = frequency to be calculated, $f0$ is the first note of the scale, and i is one of the notes of the scale. As an example, let us observe in table 1 the frequency intervals from A to A# (=Bb).

Table 1: *Frequency chart from A to Bb where f0 = 440 hertz. q* is a quarter tone.

Note	Scale
A	440 (i=0)
Aq	452.89 (i=1)
A#	466.16 (i=2)

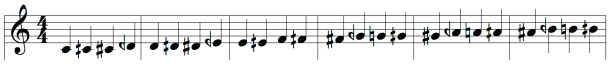

Figure 5: The scale of 24 tones in *Finale 2011*.

We created then a frequency chart using a quarter-tone scale considering all the possible quarter-tone frequencies in a piano keyboard from C0 to C8. After that, we divided the sentences in vowel-to-vowel (VV) units [cf. 14] in Praat [15] using the *BeatExtractor* plug-in [16], took the frequency mean for each VV unit using the script *f0.praat*[3], and, by comparing this VV frequency with the frequencies in our 24 tone chart, we got the musical notes with its respective octave in a piano keyboard. Of course, it was hard to find a 100% correspondence of the speech notes with the chart. That is why we devised an algorithm in *Visual Basic* within *Microsoft Excel* to automatically make this correspondence and find out what was the frequency in the chart that was closer to the mean VV frequency. For example, the VV unit "ood" in "fl**ood**ed" (see texts in [12]), recorded by subject AEF (American English female), had a mean frequency of 210 Hz, which is in-between Ab3 (207.64 Hz) and Abq[4] (213.73 Hz). By subtracting the VV frequency by the frequency to be compared (in Hz) we got that 210 Hz is closer in modulus to Ab3 (210 – 207.64 = 2.36) than Abq (210 – 213.73 = -3.73). Thus, for "ood", we found, as a result, the note Ab. We have followed this procedure for all the VV units in the corpus used by Simões and Meireles (2006) to transcribe speech with musical notation.

VV duration was extracted using the Praat script *duration.praat*. The duration chart to be compared with the VV durations was created by counting the intervals between the metronome's beat in various speeds (from 60 to 326 bpm).

[2] For a detailed account of the quarter tone scale as well as greater subdivisions of the musical scale, see Battan's thesis [13] on Alois Hába's *Neue harmonielehre des diatonischen, chromatischen, viertel-, drittel-, sechstel- und zwölftel-Tonsystems.*

[3] The Praat scripts f0.praat, duration.praat, and intensity.praat can be found at https://code.google.com/archive/p/praat-tools/downloads.

[4] Ab (=G#) ¼ tone up, which is halfway between Ab and A.

The duration of all kinds of rhythmic figures were measured. After a long process of trial-and-error, we realized that the best beat to analyze the speech data was 240 bpm. Therefore, we stuck to the measurements in this speed in order to compare the VV duration with the duration in various rhythmic figures. Exactly like we did for the frequencies, we made a formula in *Visual Basic* within *Microsoft Excel* to automatically compare the durations and find out what was the rhythmic figure in the chart that had a duration closer to the VV duration.

VV intensity was extracted using the Praat script *intensity.praat*. In this case we did not have to make any correspondence with musical descriptors, since the intensity in decibels had a scope similar to what we found for MIDI velocities (range from 0 to 127). So, we used exactly the measure intensity in the scores.

After extracting all three musical parameters (note from f0 mean, rhythmic figure from duration, and MIDI velocity from intensity in dB), we added manually these parameters in *Finale 2011*, so as to obtain the complete musical scores and corresponding audio files for all three languages.

In the process of making up the scores, we had an extra difficulty since the writing of quarter tones in musical notation software is not straightforward. To do so, we consulted the online material that explained how the notated quarter tones could sound as that (https://usermanuals.finalemusic.com/Finale2014Mac/Content/Finale/Quarter_tones.htm). In the multimedia files we have an example of a quarter-tone scale as well as the musical transcriptions for all languages.

5. Discussion

In order to exemplify the transcriptions done in this part of our research, take a look at figures 6, 7, and 8. As can be seen there, we have randomly used a 4/4 time measure, a piano score (2 clefs), and a C key for our transcriptions. Of course, this is a first step in using a quarter-tone scale for transcribing speech and thus further improvement may be necessary.

Regarding the time signature, we have to study in the future the pattern of the phrases' stress groups, so that we may divide the scores in signatures that fit the phrases strictly within the bars (see for example figure 8).

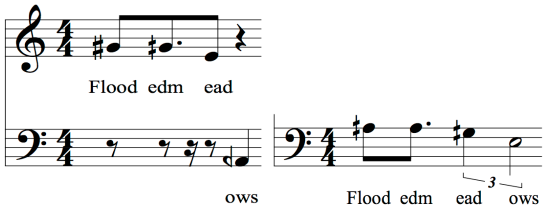

Figure 6: Musical transcription of the title of an American English poem by a female speaker (left) and a male speaker.

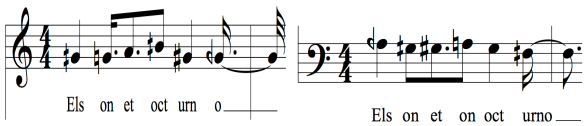

Figure 7: Musical transcription of the title of a Mexican Spanish poem by a female speaker (left) and a male speaker (right).

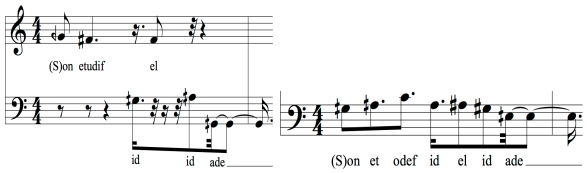

Figure 8: Musical transcription of the title of a Brazilian Portuguese poem by a female speaker (left) and a male speaker (right).

Regarding the C key, we have used it according to Schoenberg's dodecaphonic school. As no tune was more important than the other, no unique key was identified there, and, therefore, no need to identify the key. The C key was chosen just because it has no accidents, but in no ways it is to be understood as if the song was composed in C major. The same questioning is to be interpreted in our transcriptions.

Regarding the piano score, even though fretless instruments such as the violin may better be used to represent speech, we used it to facilitate the transcriptions, since in some cases we could not stick just to one single score, as can be seen in figures 6 and 8. It is important to highlight that these extreme low notes for a female voice is justified by the use of creak voice, which makes the vocal folds vibrate with a very low pattern of motion.

Finally, another problem that we had in transcribing speech was the use of voiceless vowels, which is common in Brazilian Portuguese, which results acoustically in one single VV unit, but came from 2 or more VV units. See for example, the VVs "ade(S)" and "odef" in figure 8. What happens here is that the VV duration is subjacently representing two VV units.

6. Conclusions

We presented here an innovative method of transcribing speech prosody as musical notation using a quarter-tone scale. As shown here, this method is much more reliable and precise than the technique we have used before, as suggested in section 4. There are, though, some issues that need to be taken into consideration for further improvement of the method.

First, as can be seen in figures 6 through 8, we used only a 4/4 signature for ease of transcription, and need therefore to improve the transcriptions either to include the linguistic phrases within bars or to use contemporary classical music[5] with no bars at all.

Second, we may consider not using fixed pitch representation, just the tonal patterns, as has been made by Schoenberg in *Survivor from Warsaw* (1947).

Last, we should take into consideration a better way of transcribing creak voice, since it disrupts the general musical pattern presented by the speakers, and, in a future article, discuss other issues not possible here due to space limitations.

7. Acknowledgements

The authors would like to thank the *São Paulo Research Foundation* (FAPESP grant 2015/06283-0 to the fourth author) for supporting this research, Daniel Hirst, who introduced us to the work of John Steele, and the anonymous reviewers.

[5] See for example the original version of Satie's Gnossiennes at. Freely available at http://imslp.org/wiki/Gnossiennes_(Satie,_Erik).

8. References

[1] H. S. Drinker. *Bach's Use of Slurs in Recitativo Secco*, Literary Licensing, LLC, 2013.

[2] E. Rapoport. "Schoenberg – Hartleben's Pierrot Lunaire: Speech – Poem – Melody – Vocal Performance". *J. New Music Research* 33: 71-111, 2004.

[3] E. Rapoport. "On the Origins of Schoenberg's Sprechgesang in Pierrot Lunaire." Min-Ad: Israel Studies in Musicology Online 2, 2016, at http://www.biu.ac.il/HU/mu/min-ad/06/Sprchgsng.pdf (accessed February 7, 2017).

[4] D. Deutsch, T. Henthorn, and R. Lapidis. Illusory transformation from speech to song. Journal of the Acoustical Society of America, 129, pp. 2245-2252, 2011.

[5] A. Tierney, F. Dick, D. Deutsch, and M. Sereno. "Speech versus song: Multiple pitch-sensitive areas revealed by a naturally occurring musical illusion". *Cerebral Cortex*, 2012.

[6] J. Fiehler. "The Search for Truth: Speech Melody in the Emergence of Janáček's Mature Style". In: Harmonia: Leoš Janáček: Life, Work, and Contribution. The Journal of the Graduate Association of Musicologists and Theorists at the University of North Texas, Special Issue, May 2013.

[7] J. Tyrrell. "Janáček and the Speech-Melody Myth". In: The Musical Times, Musical Times Publications Ltd., Vol. 111, No. 1530 (Aug., 1970), pp. 793-796, 1970.

[8] T. Vainiomäki. 2012. *The Musical Realism of Leoš Janáček – From Speech Melodies to a Theory of Composition*, Ph.D. dissertation, Helsinki, Finland, University of Helsinki, Faculty of Arts, 356 pages, 2012. https://helda.helsinki.fi/bitstream/handle/10138/36087/themusic.pdf?sequence. Last Visited in October 2016.

[9] J. Steele. *An essay towards establishing the melody and measure of speech, to be expressed and perpetuated by peculiar symbols*. London: Boyer and Nichols. 2nd. Edition 1779, Prosodia Rationalis. London: Nichols, 1775.

[10] V. Karbusický. "The experience of the indexical sign: Jakobson and the semiotic phonology of Leoš Janáček". *American Journal of Semiotics*, vol. 2, no 3, 35–58, 1983.

[11] P. Boulez. "Speaking, Playing, Singing". In: *Orientations: collected writings*. Faber and Faber, pp. 330–335), 1986.

[12] A. Simões and A. R. Meireles. "Speech Prosody in Musical Notation: Spanish, Portuguese and English," in *Proceedings of the 8th International Conference on Speech Prosody, Boston, USA, 2016*.

[13] S. M. Battan. *Alois Hába's Neue harmonielehre des diatonischen, chromatischen, viertel-, drittel-, sechstel- und zwölftel-Tonsystems*. Phd Thesis, The University of Rochester, New York, 1980.

[14] A. R. Meireles and V. de P. Gambarini, "Rhythm Typology of Brazilian Portuguese dialects," in *Proceedings of the 6th International Conference on Speech Prosody, Shanghai, China, 2012*.

[15] P. Boersma and D. Weenink. "Praat: Doing phonetics by computer (version 4.5.06)," http://www.praat.org/ (Last viewed December 8, 2010), 2006.

[16] P. A. Barbosa. *Incursões em torno do ritmo da fala*, Campinas: RG/Fapesp, 2006.

[17] N. von Harnoncourt. Das quasi Wort-Ton-Verhältnis in der instrumentalen Barockmusik. In: *Colloquium Music and Word. Musik und Wort*, arranged and edited by Rudolf Pečman. Colloquia on the history and theory of music at the International Musical Festival in Brno 1969. Volume 4, 225–228, 1973.