Passive voice quality analysis with melodic structure of previously recorded telephone conversations

Passive Voice-Qualitätsanalyse mit melodischer Struktur der vorher aufgezeichnete Telefonkonversation

Angel Garabitov
FDIBA an Technische Universität Sofia, Bulgarien

***Zusammenfassung* —** Der Artikel diskutiert Probleme bezüglich Werkzeuge für nicht-intrusive Bewertung der VoIP-Qualität durch eine melodische Strukturanalyse. Dies erfolgt durch eine objektive Sprachqualitätsmessung. Ziel dieser Untersuchung, die in diesem Artikel dargestellt ist, sei eine neue objektive Messmethode für VoIP-Anwendungen vorzuschlagen und zu untersuchen. Dies ist eine nicht-aufdringliche Sprachqualitäts- und Sprachverständlichkeitsmessmethode, basierend auf der Sprache zur MIDI-Umwandlung, gefolgt von einer klassischen, melodischen Strukturanalyse. Es wird ein Inhaltsanalyse-Schema beschrieben, das Aspekte der melodischen Struktur in großen Proben beurteilen kann. Der Artikel veranschaulicht die Anwendung des neuen Ansatzes für die VoIP-Anwendungen.

***Abstract* —** The paper discusses problems connected with tools to non-intrusively evaluate VoIP quality by melodic structure analysis. This is an objective voice quality measurement. The aims of the study, reported in the paper, are to investigate new objective measurement methods for VoIP applications. This is a non-intrusive speech quality and speech intelligibility measurement method, based on the speech to MIDI conversion sequitur classic melodic structure analysis. A content analytical scheme is described that can assess aspects of melodic structure in large samples. The paper illustrates the application of the new approach to the for VoIP applications.

I. EINFÜHRUNG

Die Notwendigkeit, die Sprachqualität in VoIP (Voice over IP) Anwendungen zu bewerten, ist eine wichtige Voraussetzung für technische und kommerzielle Gründe. Die vorhandenen Methoden sind möglicherweise nicht immer für VoIP-Anwendungen geeignet. Kürzlich ist die objektive Sprachqualitätsbewertung zu einem sehr aktiven Forschungsgebiet geworden. Dies ist ein Versuch, die Einschränkungen der subjektiven Tests zu umgehen, indem sie die Meinungen der menschlichen Tester algorithmisch simulieren. Diese nicht-aufdringliche Methode ist in Umgebungen wirksam, in denen das Referenz-Sprachsignal nicht zugänglich ist. Intrusive-Modelle sind zuverlässiger als die non-intrusive, da der erste Zugang zu einem Referenz-Sprachsignal hat, um das verzerrte Sprachsignal mit zu vergleichen.

Der Artikel diskutiert Probleme, die mit Werkzeugen verbunden sind, um die VoIP-Qualität durch MIDI-Analyse nicht intrusiv zu bewerten. Die Methode erkennt Beeinträchtigungen der Audioqualität für die menschliche Wahrnehmung [9]. Es ermöglicht die Qualität der VoIP-Verbindung auf einen Blick zu sehen und warnt, wenn sich die Qualität verschlechtert. Dies gibt die Möglichkeit das VoIP-Netzwerk zu beheben, noch bevor Benutzer von VoIP-spezifischen Verbindungsproblemen (Echo, Lärm oder Pausen im Gespräch) betroffen sind.

MIDI aus echtem Klang zu machen ist der Prozess, die Verwendung des Frequenzspektrums eines Klangs zu beschreiben und aufzuzeichnen. Der MIDI-Motor besteht aus einer Reihe von Bandpassfiltern, die nur die, in der Modulator-Quelle erkannten Frequenzen, erlauben. So, zum Beispiel, wenn eine menschliche Stimme verwendet wird, um einen Synthesizer-Akkord zu modulieren, wird es so klingen, als spreche ein Synthesizer - die klassische Roboterstimme aus vielen Sci-Fi-Filmen. MIDI-Datei zu erstellen bedeutet grundsätzlich die Realisierung eines Vocoders. Es ist ein Audioprozessor, der die charakteristischen Elemente eines Audiosignals erfasst und dann dieses charakteristische Signal verwendet, um andere Audiosignale zu beeinflussen. Die Technologie hinter dem Vocoder-Effekt wurde zunächst bei Versuchen zur Synthese von Sprache verwendet. Der Effekt, der Vocoding genannt wird, kann auf Datensätzen als "talking synthesizer" erkannt werden, der von Künstlern wie Stevie Wonder populär gemacht wird. Die bei der Vocoder-Analyse extrahierte Grundkomponente wird als Formant bezeichnet [10]. Der Formant beschreibt die Grundfrequenz eines Schalls und seiner zugehörigen Rauschkomponenten. Der gleiche Vorgang kann zur Analyse und Bewertung des VOIP-Signals verwendet werden. Dieser Vorgang kann die Defekte und Verzerrungen während der Übertragung des Signals erkennen. Wenn wir die menschliche Stimme als Melodie betrachten, dann werden alle Seitenverzerrungen ihre Melodie beeinflussen. Die Veränderung kann auf verschiedene Weise beurteilt werden. Eine von ihnen ist die Auswertung der melodischen Erwartung. Wenn jemand eine Melodie hört, ist es üblich eine Erwartung zu haben, was als nächstes kommen wird. Das alles gilt für die Rede. Eine wichtige Studie zu diesem Thema ist das Implikation-Realisierungsmodell von Narmour [7], das die melodische Erwartung vernünftigerweise vorhersagen kann. Hier werden einige Hauptpunkte erwähnt. Das erste Prinzip des Modells ist eine Registernummer und es bedeutet, dass kleine Intervalle dazu neigen, eine Fortsetzung der Tonhöhenrichtung zu verfolgen, während große Intervalle eine Richtungsänderung erwarten. Intervalldifferenz ist das Prinzip, dass kleine Intervalle andere Intervalle von ähnlicher

Größe implizieren, während große Intervalle kleinere Intervalle implizieren. Wenn ein zweites Intervall zu einer Tonhöhe führt, die der ursprünglichen Tonhöhe nahekommt, haben wir das Prinzip der Registrierungsrendite. So erwarten die Zuhörer, dass die Sprünge auf eine ähnliche Tonhöhe zurückkehren. Das Prinzip der Nähe definiert einfach, dass kleine Intervalle eher erwartet werden als große Intervalle. Das letzte Prinzip ist die Schließung und es tritt auf, wenn die Melodie die Richtung ändert oder wenn ein großes Intervall von einem kleineren Intervall gefolgt wird. Das Modell wurde von Frau Carol L. Krumhansl [1] quantifiziert. In ihrer Arbeit schlägt sie auch ein neues Prinzip namens Konsonanz vor. Dieses Prinzip besagt, dass Unisionen, perfekte Viertel, Fünftel und Oktaven bevorzugte Intervalle sind. Dies ist jedoch eine sehr kontextspezifische Behauptung, da Melodien je nach Genre viele weitere chromatische und dissonante Intervalle haben können.

Sprachauswertungssysteme sind Systeme, die ein akustisches Signal durch Algorithmen analysieren. Diese Algorithmen basieren auf einer Anzahl von Theorien, die vorschlagen, welche Merkmale des Sprachsignals in einer gegebenen Sprache einen Klang erzeugen. Mathematische Methoden bestimmen die wichtigen Parameter des akustischen Signals und verwandeln sie in eine unterschiedliche Vollständigkeit in der gewünschten Form [6].

Es gibt Möglichkeiten, diese Erwartungen in einer Evaluierungsfunktion zu nutzen. Die Modelle sind rechenintensiv, da sie ihre Ergebnisse auf die Zeit- und / oder Frequenzbereichsanalyse des zu prüfenden Sprachsignals stützen. Sie verlangen auch, dass der Testaufruf für eine beträchtliche Dauer aufgezeichnet wird, bevor er analysiert werden kann. Daher eignen sie sich nicht für die Echtzeit- und kontinuierliche Überwachung der Sprachqualität, sind aber für die anschließende Analyse anwendbar.

II. Vorgeschlagenes Algorithmus für VOIP Qualität durch Midianalyse

Die Aufgabe besteht darin, das Sprachsignal zu studieren, Parameter zu identifizieren und zu visualisieren, die Klänge zu erzeugen (Phoneme), diese Parameter messen und klassifizieren.

Die Studie untersuchte folgende Aspekte:

- Signalverarbeitung;
- Hörmuster;
- Artikulationsmodelle;
- Aussprachemuster;
- Suchalgorithmen;
- Lernalgorithmen;

Die möglichen Optionen, Metadaten aus dem Sprachsignal abzurufen, können mit denen verglichen werden, die die Qualität einer musikalischen Phrase auswerten [2]. Ähnliche musikalische Analyse-Aufgaben wurden weitgehend entwickelt und hier ist ein Versuch, sie auf die Klanganalyse anzuwenden. Die musikalische Analyse selbst impliziert die erste Umwandlung in ein entsprechendes Format.

Umwandlung in MIDI

Ein Sprachcodierungssystem zum Codieren eines digitalisierten Sprachsignals in ein Standard-Digitalformat, wie beispielsweise MIDI. Das Pitch-Detection-System ist gut, aber nicht gut genug bei komplexen oder harmonisch reichen Inhalt. Was es tut, ist die Analyse und Annäherung der Häufigkeit dessen, was er summt und in eine Folge von MIDI-Noten umwandelt.

Die stimmhaften Töne resultieren aus den Vokal-Akkorden, die vibrieren und so den Luftstrom aus den Lungen unterbrechen und einen Frequenzbereich von Klängen von etwa 50 bis zu 500 Hz erzeugen [5].

Plosivgeräusche sind plötzliche Luftstöße, die zum Beispiel ausgeschaltet werden, wenn der Vokaltrakt geschlossen ist und plötzlich freigegeben wird oder der Mund plötzlich geöffnet wird. Alle diese Töne werden durch die Sünden und Nasenhöhlen der Person beeinflusst und alle machen das, was wir als normale menschliche Sprache verstehen.

1) Grundlagen Klangbeispiele und Entschlüsselung der MIDI-Datei

Die verwendete Probe ist von Open Speech Repository [11] heruntergeladen. Die Aufnahmebedingungen sind Paketverlust 50% und Korrelation 15%. Für die Erstellung des Fehlersignals wird der Befehl Linux tc verwendet. In Abbildung 1 ist der Vergleich des guten und schlechten Signals dargestellt.

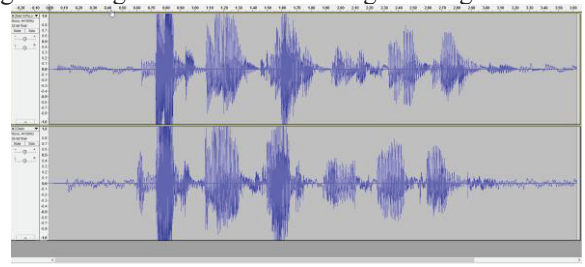

Abb. 1. Gutes und schlechtes Signal

2) MIDI-Extraktion mit dem Signal

In Tabelle I ist der Vergleich des guten und schlechten Signals in MIDI Format dargestellt.

TABELLE I. MIDI Format

Gutes Signal	***Schlechtes Signal***
MFile 1 2 120 MTrk 0 Meta Copyright "Created by TS-AudioToMIDI 3.30" 0 TimeSig 4/4 24 8 0 KeySig 0 major 0 Tempo 500000 0 Meta TrkEnd TrkEnd MTrk 22 PrCh ch=1 p=0 22 On ch=1 n=39 v=23 58 Off ch=1 n=39 v=0 75 On ch=1 n=38 v=18 … 876 Meta TrkEnd TrkEnd	MFile 1 2 120 MTrk 0 Meta Copyright "Created by TS-AudioToMIDI 3.30" 0 TimeSig 4/4 24 8 0 KeySig 0 major 0 Tempo 500000 0 Meta TrkEnd TrkEnd MTrk 73 PrCh ch=1 p=0 73 On ch=1 n=24 v=20 85 Off ch=1 n=24 v=0 86 On ch=1 n=34 v=24 … 813 Meta TrkEnd TrkEnd

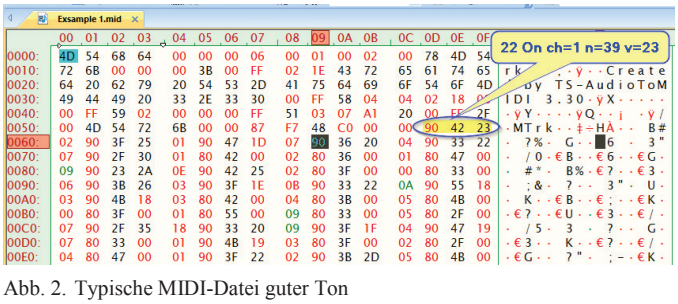

Abb. 2. Typische MIDI-Datei guter Ton

In Abbildung 2 ist ein MIDI Dateien in einen Hexformat dargestellt

3) Konvertieren von MIDI-Dateien in Matlab

Die grundlegenden Funktionen in MIDI Toolbox lesen und manipulieren Typ 0 und Typ 1 MIDI-Dateien. Nicht Ematrix (oder cmat) bezieht sich auf eine Matrixdarstellung von Notenereignissen in einer MIDI-Datei.

TABELLE II. MIDI Kodierung

ONSET (BEATS)	*DURAT. (BEATS)*	*MIDI channel*	*MIDI pitch*	*VELO-CITY*	*ONSET (SEC)*	*DURAT. (SEC)*
0.6083	0.1000	0	24.0	20.0	0.3042	0.0500
0.7167	0.1417	0	34.0	24.0	0.3583	0.0708
0.8583	0.1583	0	32.0	21.0	0.4292	0.0792
1.1583	0.2000	0	34.0	42.0	0.5792	0.1000
1.6750	0.1750	0	48.0	49.0	0.8375	0.0875
1.8417	0.1000	0	36.0	26.0	0.9208	0.0500

Die erste Spalte zeigt den Beginn der Noten in Beats an (auf der Grundlage von Zecken pro Viertelnote) und der zweiten Spalte die Dauer der Noten in denselben Beat-Werten. Die dritte Spalte bezeichnet den MIDI-Kanal (1-16) und die vierte MIDI-Tonhöhe, wobei der mittlere C (C4) 60 ist. Die fünfte Spalte ist die Geschwindigkeit, die beschreibt, wie schnell die Taste der Note gedrückt wird, mit anderen Worten, wie laut die Note gespielt wird (0-127). Die letzten beiden Spalten entsprechen den ersten beiden (Beginn in Beats, Dauer in Beats), außer dass Sekunden anstelle von Beats verwendet werden.

4) Statistische Auswertung der Dateien in Matlab

Zuerst können wir die Notenverteilung der Klangbeispiele untersuchen.

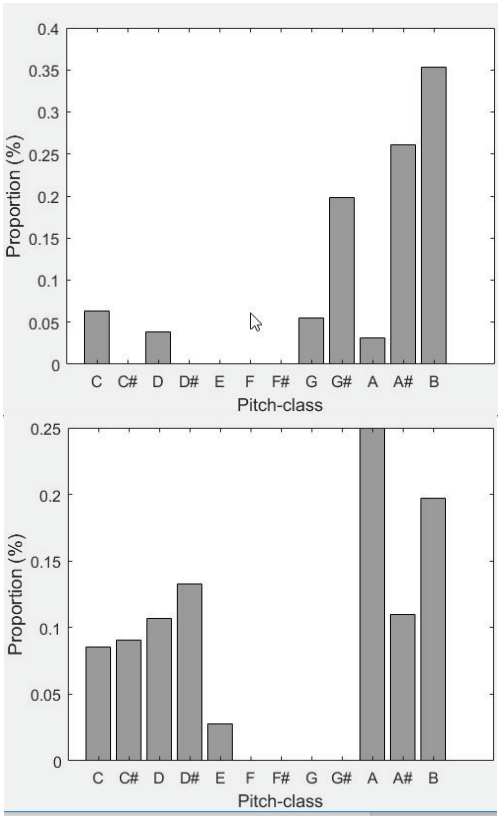

Abb. 3. Notizverteilung der Klangbeispiele

Der Anteil der Zwei-Noten-Fortsetzungen - Anteil der Tonübergänge

Die melodische Kontur beschreibt die Gesamtform der Klangprobe. Die durchgezogene Linie stellt eine gröbere melodische Kontur dar, die nützlicher sein könnte, um die Gesamtstruktur der Tonprobe herauszufinden.

Diese Visualisierung der Tonalität kann verwendet werden, um die Schwankungen des Schlüsselzentrums und der Schlüsselstärke im Laufe der Zeit zu zeigen.

Diese Visualisierung der Tonalität kann verwendet werden, um die Schwankungen des Schlüsselzentrums und der Schlüsselstärke im Laufe der Zeit zu zeigen.

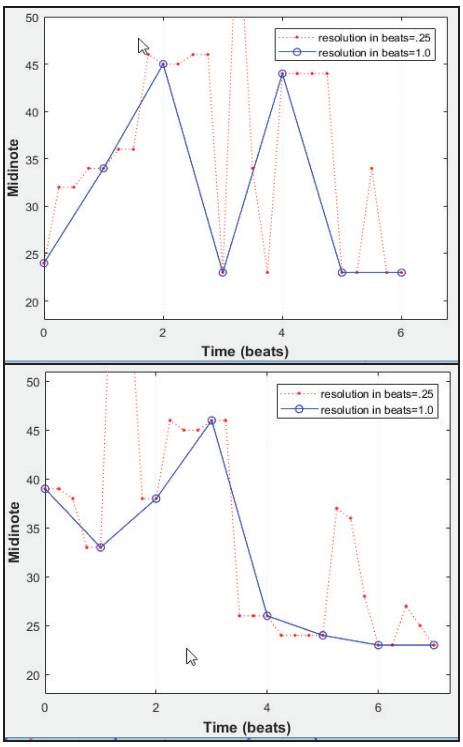

Abb. 4. *Gesamtstruktur der Tonprobe*

5) Melodische Erwartungen und Sprachverständlichkeit [3].

Die früheren Arbeiten zur melodischen Erwartung haben gezeigt, wie sich die Musik auf die üblichen psychologischen Erwartungsprinzipien stützt, die in kognitiv orientiertem musiktheoretischen Modell von Narmours [7] gefangen wurden. Das Modell stützt sich auf die Gestalt-basierten Prinzipien der Nähe, der Ähnlichkeit und der guten Fortsetzung und wurde gefunden, um die melodischen Erwartungen der Zuhörer recht gut vorhersagen zu können. Das Modell arbeitet mit Blick auf implizite Intervalle und realisierte Intervalle. Der ehemalige schafft Implikationen für die Fortsetzung der Melodie und das nächste Intervall führt seine Implikationen aus.

6) Melodische Komplexität und Sprach-verständlichkeit [3].

Gelegentlich ist es interessant zu wissen, wie kompliziert, schwierig oder "originell" eine Melodie der Rede ist. Diese Beziehung ist in Form einer invertierten U-Funktion, wo die populärsten Themen von mittlerer Originalität sind. Infolgedessen sind die einfachsten Themen nicht populär (sie können als "banal" angesehen werden) und auch nicht die komplexesten. Es gibt auch andere Verwendungen für eine melodische Komplexitätsmaßnahme, wie sie es als Hilfe bei der Klassifizierung von melodischem Material zu verwenden (Toiviainen & Eerola, [3]). Simontons Modell der melodischen Originalität basiert auf Tonübergangswahrscheinlichkeiten. Die Ausgabe dieses Modells (compltrans) erzeugt eine Umkehrung der gemittelten Wahrscheinlichkeit, skaliert zwischen 0 und 10, wo ein höherer Wert eine höhere melodische Originalität anzeigt.

Eine weitere Möglichkeit, die melodische Komplexität zu beurteilen, besteht darin, sich auf die Klang- und Akzentkohärenz zu konzentrieren, und auf die Höhe der Pitch-Skips und die Kontur-Selbstähnlichkeit zeigt sich die Melodie [4]. Eine alternative Maßnahme der melodischen Komplexität ist in der kontinuierlichen Messung der Notenereignisverteilung (Pitch-Class, Interval) Entropie (Use Mov-Fenster und Entropie und verschiedene Verteilungsfunktionen) verankert. Diese Maßnahme erzeugt für jeden Punkt in der Melodie melodische Vorhersagungswerte (daher der Begriff kontinuierlich). Es wurde festgestellt, dass diese Werte den Vorhersehbarkeitsangaben entsprechen, die von den Zuhörern in Experimenten gegeben wurden. Diese Maßnahme bietet die Möglichkeit, die Moment-in-Moment-Schwankungen der melodischen Vorhersagbarkeit zu beobachten.

Viele Software- und Hardwarelösungen wurden implementiert.

III. Berechnen der melodischen Funktionen

Es gibt viele Möglichkeiten, um die Melodie zu bestimmen, die zu unserem Zweck anwendbar ist. Im Wesentlichen sind sie sehr komplex zu beschreiben. Die einfachste Methode zur Anwendung ist die Methode von Euler (Grad der Melodie [8]).

Euler schlug vor, dass der Grad der Melodie von den Berechnungen des Geistes abhängt: je weniger die Berechnungen, desto angenehmer die Erfahrung.

Eine geringe Anzahl von Berechnungen führt zu einem hohen Wert für die Melodie, während eine hohe Anzahl von Berechnungen einen niedrigen Wert ergibt.

Dieses Prinzip wird durch eine numerische Technik umgesetzt, die auf der Zerlegung der natürlichen Zahlen in ein Produkt von Kräften verschiedener Primzahlen basiert.

Wenn p_i ..., p_n sind verschiedene Primzahlen und e_1, ..., e_n sind verschiedene Kräfte, dann kann jede natürliche Zahl a ausgedrückt werden als

$$a = p_1^{e_1} p_1^{e_2} \cdots p_1^{e_n} = \prod_{i=1}^{n} p_i^{e_i} \qquad (1)$$

$$G(a) = 1 + \sum_{k=1}^{n} e_k (p_k - 1) \qquad (2)$$

```
function y = gradus(nmat)
if isempty(nmat), return; end
if ~ismonophonic(nmat);
return; end
int=abs(diff(pitch(nmat)));
int=mod(int,12)+1;
frq1=[1,16,9,6,5,4,45,3,8,5,16,15];
frq2=[1,15,8,5,4,3,32,2,5,3,9,8];
for k=1:length(int)
        n(k)=frq1(int(k));
        d(k)=frq2(int(k));
end
for s=1:length(n)
        gs(s)=suavitatis(n(s),d(s));
end
y=mean(gs);
        g=n*d;
        g2=factor(g);
        for i=1:length(g2)
                g3(i)=1*g2(i)-1;
        end
        g=sum(g3)+1;
```

Abb. 5. Algorithm in Matlab

Euler benutzt Primzahlen in der Bemühung, die Melodie, den "Grad der Süße" zu quantifizieren - oder wie er es die "gradus suavitatis" - der Töne nannte. Der Gradus suavitatis einer einzigen Note wurde als 1 und darüber hinaus genommen, wenn das Frequenzverhältnis von zwei Noten ***m: n*** ist und wenn das kleinste gemeinsame Vielfache von ***m*** und ***n*** ***L*** ist, hat er die Definition gemacht.

$$G(m,n) = 1 + \prod_{\substack{p\ prime \\ p\ divides\ L}} (p-1) \qquad (3)$$

Das Algorithm von Euler ist in Matlab in realisier. (Abbildung 5)

IV. Schlussfolgerungen und zukünftige Arbeiten

Das Problem der Echtzeit-Schätzung von VoIP ist von erheblichem Interesse. Dieses Papier hat einen Ansatz zur Lösung dieses Problems gezeigt, indem er die Melodien des Signals einsetzt. One of the main objectives of this research was to estimate the effect of fragmentation on speech quality. Eines der Hauptziele dieser Forschung war es, die Wirkung der Fragmentierung auf die Sprachqualität abzuschätzen.

Der Fokus der aktuellen Forschung liegt auf der Schätzung der Wirkung aller VoIP-Verkehrsparameter, die die Hörqualität eines Telefonanrufs im Mähdrescher beeinflussen. Ein zukünftiges Ziel wäre es, ein neuronales Netzwerkmodell für die Abbrsation-Qualitätsschätzung eines Aufrufs abzuleiten. Die Konversationsqualität leidet unter der Zunahme der End-to-End-Verzögerung eines Anrufs. Klar wäre das nächste Ziel, die besondere Wirkung von VoIP-Verkehrsparametern und deren Auswirkungen auf die Signalqualität abzuschätzen.

Die digitale Signalmelodie-Bewertung wird verwendet, um ein neuronales Netzwerk zu trainieren, um die Klangqualität von VoIP zu beurteilen.

Danksagung

Diese Publikation wurde finanziert durch Vertrag zur Doktorandenunterstützung mit Nummer 172PD0010-07 der TU – Sofia.

Literaturverzeichniss

[1] Carol L. Krumhansl, *Cognitive Foundations of Musical Pitch*. Oxford: Oxford University Press, 1990, 307pp, ISBN 0-19-505475-X.

[2] David Huron and Matthew Royalp "What Is Melodic Accent? Converging Evidence from Musical Practice", Music Perception: An Interdisciplinary Journal, Vol. 13, No. 4 (Summer, 1996), pp. 489-516 Published by: University of California Press Stable, URL: http://www.jstor.org/stable/40285700

[3] Eerola, T. & Toiviainen, P. *MIDI Toolbox: MATLAB Tools for Music Research*. University of Jyväskylä: Kopijyvä, Jyväskylä, Finland, (2004).

[4] Marc Lema, "Music and Schema Theory: Cognitive Foundations of Systematic Musicology", Springer Science & Business Media, Dec 6, 2012 - Science - 234 pages

[5] Marcel Wältermann, *Dimension-based Quality Modeling of Transmitted Speech*, Springer-Verlag Berlin Heidelberg 2013.

[6] Moller, S., "Assessment and Prediction of Speech Quality in Telecommunications", Kluwer Academic Publishers, Boston/Dordrecht/London, 2000, 116-117.

[7] Narmour, E. *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*. Chicago: University of Chicago Press, 1990, ISBN 0-226-56845-8.

[8] Patrice Bailhache, *Euler and music*, Nantes, 1997

[9] Patrik N. Juslin, John Sloboda, *Handbook of Music and Emotion: Theory, Research, Applications*, Oxford University Press, Mar 17, 2011 - Psychology - 992 pages

[10] Trevor A. Harley, *Speech Perception and Spoken Word Recognition, Current Issues in the Psychology of Language Series*, first published 2017 by Routledge.

[11] http://www.voiptroubleshooter.com/open_speech/american.html.