Using Formant as an indication of speech intelligibility in VOIP

Verwenden der Formant als Indikation der Sprachverständlichkeit in VOIP

Angel Garabitov*, Aleksandar Tsenov†

Fakultät für deutsche Ingenieur- und Betriebswirtschaftsausbildung, TU – Sofia
Sofia, Bulgaria, *angel.garabitov@fdiba.tu-sofia.bg, †aleksandar.tsenov@fdiba.tu-sofia.bg

***Abstract* —** Voice over Internet Protocol (VoIP) is a revolutionary technology which is acting as a platform for the development of latest trends in modern communication world. The speech signal quality in VoIP is governed by the speech coding techniques employed. Conveying the message is one of the most important goals. This can be challenging since the intelligibility of the speech may be harmed at various stages before, during and after the transmission process from sender to receiver. Effective quality evaluation methodologies are important for system development and refinement, particularly by adopting user feedback-based measurement. The article researches an affect-based QoE (Quality of Experience or QoE-based) evaluation framework which attempts to capture users' perception while they are engaged in voice communication. The article makes a proposal to use a formant as criteria for the prediction of subjective voice quality assessment. In the work we propose a new measure which show good prediction results with the intelligibility of nonlinear processed speech. The newly proposed measure is of a low computational complexity and mathematically tractable which make them suitable for optimization of new signal processing solutions which aim for improving speech intelligibility.

***Zusammenfassung* —** Voice over Internet Protocol (VoIP) ist eine revolutionäre Technologie, die als Plattform für die Entwicklung der neuesten Trends in der modernen Kommunikationswelt dient. Die Sprachsignalqualität in VoIP wird durch die verwendeten Sprachcodierungstechniken bestimmt. Die Übermittlung der Botschaft ist eines der wichtigsten Ziele. Dies kann eine Herausforderung darstellen, da die Verständlichkeit der Sprache in verschiedenen Stadien vor, während und nach dem Übertragungsprozess vom Sender zum Empfänger beeinträchtigt sein kann. Effektive Qualitätsbewertungsmethoden sind wichtig für die Systementwicklung und -verfeinerung, insbesondere durch die Verwendung von Benutzer-Feedback-basierten Messungen. Der Artikel untersucht ein affektives QoE (Quality of Experience oder QoE-basiertes) Evaluierungs-Framework, das versucht, die Wahrnehmung der Nutzer während der Sprachkommunikation zu erfassen. Der Artikel schlägt vor, einen Formanten als Kriterium für die Vorhersage der subjektiven Sprachqualitätsbewertung zu verwenden. In der Arbeit schlagen wir ein neues Maß vor, das gute Vorhersageergebnisse mit der Verständlichkeit von nichtlinear verarbeiteter Sprache zeigt. Die neue vorgeschlagene Maßnahme ist von geringer rechnerischer Komplexität und mathematisch handhabbar, was sie zur Optimierung neuer Signalverarbeitungslösungen geeignet macht, die auf eine Verbesserung der Sprachverständlichkeit abzielen.

I. EINFÜHRUNG

Sogar in einer reinen Klangsilbe kann es eine ganze Reihe von Klangfrequenzen sein. Die menschlichen Stimmorgane fügen dem Grundton, den die Stimmbänder bilden, zusätzliche harmonische Komponenten hinzu. Diese Komponenten verleihen der Stimme Farbe, wodurch wir insbesondere die Sprache der Personen, die wir kennen, erkennen können. Als Ergebnis der Forschungen wurde gefunden, dass vier Frequenzen, die sich aktiv in den Resonanzhohlräumen des Stimmtrakts bilden, aktiv an der Sprachbildung beteiligt sind. Diese Frequenzen werden Formanten genannt. Bei der Artikulation können sich sowohl die Amplitude als auch die Frequenz der Formatkomponenten des Klangs ändern. Gleichzeitig bleibt jedoch die Anzahl der Formanten in Sprachtönen konstant und beträgt immer vier. [1]

Bei Geräuschen ist es schwierig, Formantkomponenten darin zu identifizieren.

In Bezug auf die Wissenschaft, die sich mit der Erkennung und Synthese von Sprache befasst, stellen wir hier ein anderes Konzept der Einheit des Tonsystems vor – das Phonem.

Phoneme werden nicht direkt mit Wörtern oder Silben identifiziert. Sie spielen die Rolle unteilbarer Teilchen, Zungenatome und sind Tonfolgen. [2,3] Alle anderen Sprachkonstrukte wie Silben und Wörter setzen sich aus Phonemen zusammen. Phoneme verfügen über zahlreiche Funktionen, die zur Klassifizierung und Erkennung verwendet werden können.

Das Wissen, dass die Frequenzen der ersten Formanten der Phoneme (d. H. Die niedrigsten akustischen Resonanzen des Stimmtrakts) primäre Informationsträger in der menschlichen Sprache sind.

Dies hat viele Forscher veranlasst, Mittel zum Abschätzen von Formantenfrequenzen aus dem akustischen Signal zu entwickeln. Ein Algorithmus zur automatischen Formantenschätzung kann Anwendungen finden, darunter: Klassifizierung von vokalähnlichen phonetischen Einheiten in einem sprecherunabhängigen Erkennungssystem [4], Erfassung von Parametern für Diphone-Text-zu-Sprache-Systeme [5] und Datenreduktion für die Forschung in der akustischen Phonetik.

Die Natur der Formant - Stimmfalten (Bündel) schwanken, wenn sie unter dem Bernoulli-Effekt durch sie geblasen werden.

Anregung einer Volumengeschwindigkeit mit einer einzigen Amplitude am Eingang des Wellenleiters (1).

$$u(t)=sin\ (\omega\ t) \tag{1}$$

Die Übertragungsfunktion ist in der Form:

$$u(x,t)=\frac{\sin(\omega t)\cos\left(\frac{\omega(l-x)}{c}\right)}{\cos\left(\frac{\omega l}{c}\right)} \tag{2}$$

Maxima im Spektrum entsprechen stehenden Wellen mit einer Wellenlänge (3):

$$\lambda(n)=\frac{1}{1/4+n/2}, n=0,1,\ldots \tag{3}$$

Oder Frequenzen:

$$F(n)=\frac{c}{4l}(1+2n), n=0,1,\ldots \tag{4}$$

Die Übertragungsfunktion des Vokaltrakts und folglich das Sprachsignal sind durch Maxima in dem um mehrere hundert Hertz getrennten Spektrum gekennzeichnet und hängen im Wesentlichen von der Form des Vokaltrakts ab. Für Vokale werden diese Maxima Formanten genannt.

Eine effiziente und kompakte Darstellung der zeitabhängigen Sprachmerkmale bietet potenzielle Vorteile für die Spracherkennung. Formant-Tracking-Methoden, die auf Linear Prediction Analysis (LPC) basieren, haben beträchtliche Aufmerksamkeit gefunden. Wurzelfindungsalgorithmen werden verwendet, um die Nullstellen des LPC-Polynoms zu finden, oder lokale Maxima der LPC-Hüllkurve [6] werden unter Verwendung von Spitzenpicking-Techniken gesucht. Das Problem bei Verfahren zur numerischen Lösung ist, dass die Bestimmung der Formantenfrequenzen und -bandbreiten nur für komplex konjugierte Pole und nicht für echte Pole erfolgreich ist.

Es kann jedoch bestimmte Aspekte geben, aufgrund der formantbasierten Parameter attraktiv sind:

- Formanten gelten als robust gegenüber Kanalverzerrungen und Rauschen.
- Formant-Parameter können ein Mittel sein, um das Problem einer Nichtübereinstimmung zwischen Trainings- und Testbedingungen zu lösen.
- Es gibt eine enge Beziehung zwischen Formant-Parametern und modellbasierten Ansätzen der Sprachwahrnehmung und -produktion.

II. LPC-Formant-Schätzung

Das Sprachsignal wird durch die Faltung der Anregungsquelle und der zeitveränderlichen Komponenten des Vokaltrakt-systems erzeugt. [7] Diese Anregungs- und Vokaltraktkomponenten sind von dem verfügbaren Sprachsignal zu trennen, um diese Komponenten unabhängig zu untersuchen. Zur Dekonvolution der gegebenen Sprache in Anregungs- und Vokaltraktsystemkomponenten werden Methoden entwickelt, die auf homomorpher Analyse wie Cepstralanalyse basieren. Da die Cepstralanalyse die Dekonvolution von Sprache in Quell- und Systemkomponenten durchläuft, indem sie durch den Frequenzbereich wandert, wird die Dekonvolutionsaufgabe ein rechenintensiver Prozess. Um eine solche Art von Berechnungskomplexität zu reduzieren und die Quellen- und Systemkomponenten aus der Zeitdomäne selbst zu finden, wird die lineare Vorhersageanalyse entwickelt.

Die isolierten Signale werden mit einem linearen prädiktiven Codierungsmodell analysiert, um die Formanten zu finden.

Beschreibung des Algorithmus:

1. Führen Sie ein autoregressives Autokorrelationsmodell (LPC) der Ordnung LPC_COEFF aus
2. Berechnen Sie die komplexen Wurzeln des LPC-Modells
3. Ignoriere Wurzeln mit positivem Imaginärteil (beliebig, hätte die Negative auch ignorieren können)
4. Konvertieren Sie die Wurzeln in den Frequenzbereich basierend auf der Abtastfrequenz

Die Formantenfrequenzen (Abbildung 1) des normalen Schreiens von Säuglingen wurden unter Verwendung von drei verschiedenen Schätztechniken gemessen; Schallspektrographie, lineare prädiktive Codierung (LPC) und Leistungsspektrumanalyse.

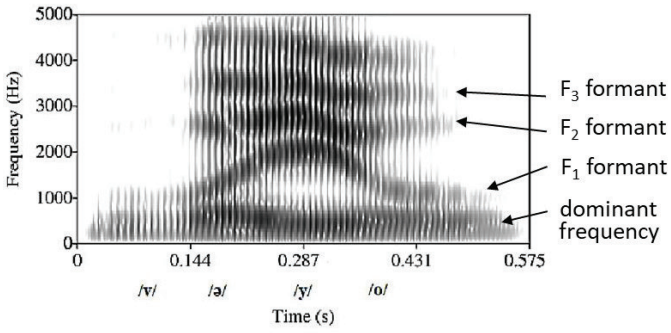

Abb. 1. Beispiel für Formanten

Resonanzen in Sprache Vocal Tract (Hals + Zunge + Lippen) agiert als variabler Resonator. Resonanzen sind Formanten. [8]

III. Spracherkennung der deutschen langen Monophthonge mit Formanten

Es gibt sieben lange Monophthonge in deutscher Sprache (Tabelle 1).

TABELLE I. Lange Monophthonge in deutscher Sprache

Lange Monophthonge	Geschrieben	Gesprochen
i:	viel	/f***i:***l/
e:	Beet	/b***e:***t/
a:	Saat	/z***a:***t/
o:	Boot	/b***o:***t/
u:	Hut	/h***u:***t/
y:	Rübe	/r***y:***bə/
ø:	Öl	/***ø:***l/

Die verwendeten Vokale stammen von den E-Learning-Einheiten auf dem Virtual Linguistics Campus. [9]

Auf Abbildung 2 sind sichtbar: die Amplitude, der Frequenzgang, erste und zweite Formant. Sie sind durch die Algorithm der Praat Programme [10] berechnet.

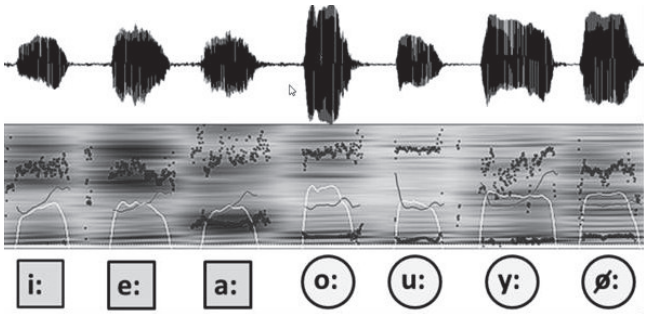

Abb. 2. Lange Monophthonge mit Formanten

IV. Berechnungen für Monophthonge I

Bevor wir die Spracherkennung versuchen, müssen wir das Sprachsignal vorverarbeiten. Während dieser Verarbeitung sollten Rausch- und Fremdsignale entfernt werden, deren Frequenzspektrum außerhalb des menschlichen Sprachspektrums liegt.

Um das Problem der Erkennung zu lösen, müssen die Hauptmerkmale der Sprache identifiziert werden, die in den nachfolgenden Stufen des Erkennungsprozesses verwendet werden. Primärmerkmale werden durch Analyse der spektralen und dynamischen Eigenschaften des Sprachsignals unterschieden.

Vokale können phonetisch schwer zu beschreiben sein, da sie Punkte oder Bereiche innerhalb eines zusammenhängenden Raums sind. Die ersten beiden Formanten sind wichtig für die Bestimmung der Qualität von Vokalen. Zur Hervorhebung der informativen Merkmale des Sprachsignals wird die spektrale Darstellung von Sprache verwendet.

Wir berechnen die Frequenzantwort des Tons i: nach den klassischen Methoden. Das Ergebnis ist auf Abbildung 3 dargestellt. Im Frequenzgang sind die schwarzen Linien der Formate deutlich sichtbar.

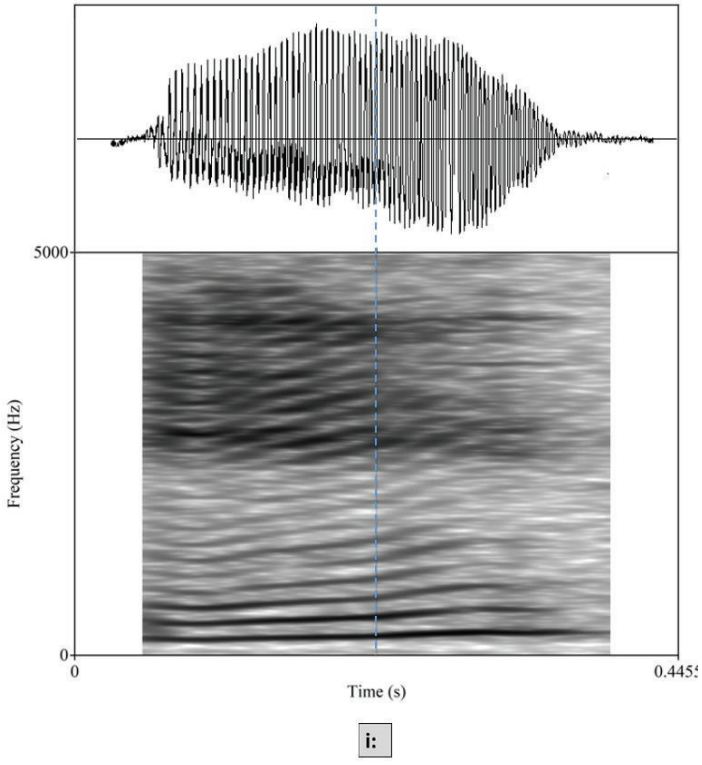

Abb. 3. Amplitude und Frequenzgang der Monophthong i:

Abbildung 4 zeigt eine Frequenzantwort mit der Basisfrequenz beider Formate für einen bestimmten Zeitpunkt.

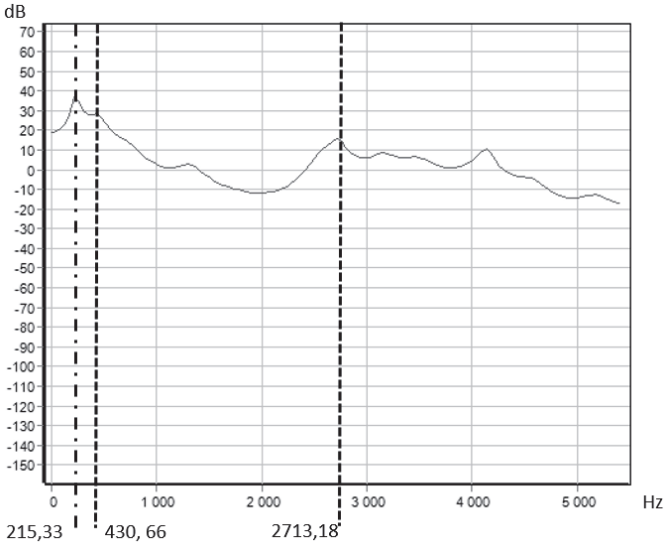

Abb. 4. Frequenzband der Monophthong i

Mit dem in Punkt 2 beschriebenen LPC-Formant-Schätzung-Algorithmus werden die Werte der ersten beiden Formate für das gesamte Zeitintervall berechnet.

Wir haben MATLAB verwendet, um einen Algorithmus zum Schätzen der Positionen der Formanten von stimmhaften Sprachintervallen basierend auf den Positionen der Sprachpolynomwurzeln zu realisieren, wie sie aus einer rahmenbasierten Analyse eines Sprachsignals unter Verwendung des Verfahrens der linearen Vorhersage-Analyse erhalten werden. Ein Schlüsselaspekt dieses Algorithmus besteht darin, Regionen von quasi zusammenhängenden Sprachrahmen zu finden (wobei die ersten drei Formanten über die Dauer der stimmhaften Region eine zusammenhängende Region bilden) und dann diese Regionen (sowohl rückwärts als auch vorwärts) basierend auf einer etwas schwächeren Region zu erweitern Kriterium für das zusammenhängende Verhalten der Formanten.

Beide Formanten sind auf Abbildung 5 dargestellt.

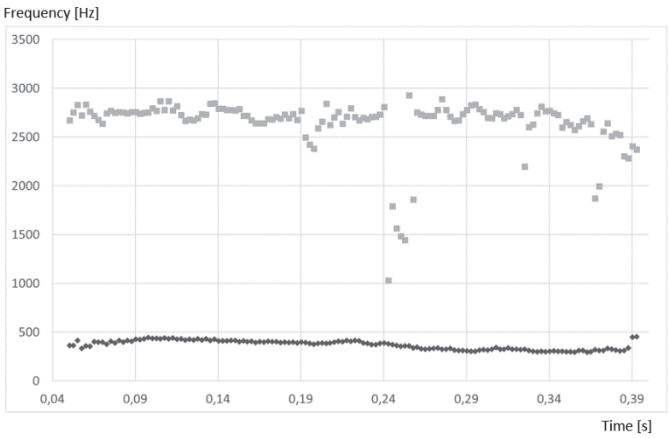

Abb. 5. Erste und zweite Formant der Monophthong i:

V. Schlussfolgerung

Hier haben wir versucht, den akustischen Raum in Tabelle 2 zu bringen. Die ersten beiden Formanten sind für die Bestimmung der Vokalqualität am wichtigsten, und dies wird häufig in Form einer Auftragung der ersten Formanten gegen die zweiten Formanten angezeigt.

Die Erkennung kontinuierlicher Sprache ist ein mehrstufiger Prozess. Nach der Vorverarbeitung des Sprachsignals und dem

Extrahieren von informativen Merkmalen daraus werden die lexikalischen Elemente der Sprache extrahiert. Dies ist die erste Stufe der Anerkennung.

TABELLE II. VERHÄLTNIS DER FORMANTEN DER LANGE MONOPHTHONGE IN DEUTSCHER SPRACHE

Monophthonge	F2/F1
i:	9,47
e:	1,43
a:	3,62
o:	8,05
u:	11,72
y:	12,19
ø:	6,23

Wenn Sie mit einem Lehrer ein Netzwerk unterrichten, können Sie dem Netzwerk beibringen, Objekte zu erkennen, die zu einer vorgegebenen Gruppe von Klassen gehören. Wenn das Netzwerk ohne Lehrer trainiert wird, können Objekte nach ihren digitalen Parametern in Klassen eingeteilt werden.

Der Ansatz basiert auf der Auswahl lexikalischer Elemente (Phoneme, Allophone, Morpheme usw.) in der Sprache.

Das hier erwähnte Verfahren zur Isolierung der lexikalischen Elemente der Sprache basiert auf der Verwendung eines Verhältnisses des zweiten zu dem ersten Format.

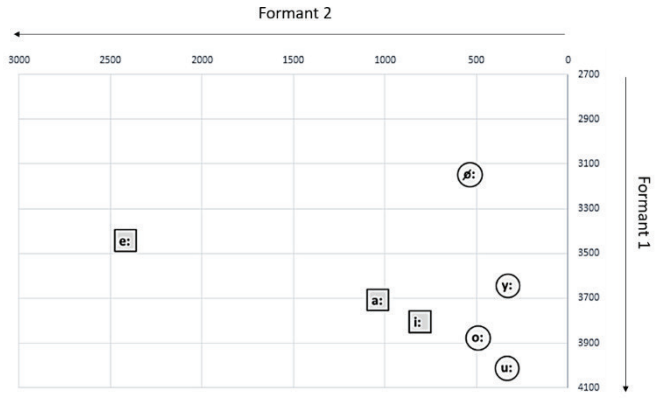

Abb. 6. Grafische Darstellung der Formanten der 7 langen Monophthonge

ANERKENNUNG

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

LITERATURVERZEICHNISS

[1] Rabiner L., R. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978

[2] Rishma Rodrigo R., R. Radhika, M. Vanitha Lakshmi, „Speech Enhancement of Vowels Based on Pitch and Formant Frequency“, World Academy of Science, Engineering and Technology, *International Journal of Computer and Systems Engineering*, Vol:9, No:3, 2015

[3] Madisetti V., *Video, Speech, and Audio Signal Processing and Associated Standards (The Digital Signal Processing Handbook, Second Edition)*, 2009

[4] Prodeus A., „Formant-Modulation Method of Speech Intelligibility Evaluation“ *Measuring and Exactness,* MEMSTECH'2011

[5] Gopi E. S., *Digital Speech Processing Using Matlab*, Springer, 2014

[6] Furui S., *Digital Speech Processing, Synthesis and Recognition*, CRC Press, 2000

[7] Zheng-Hua Tan, B. Lindberg, *Automatic speech recognition on mobile devices and over communication networks*, Springer, 2008

[8] Dong Yu, Li Deng, *Automatic Speech Recognition A Deep Learning Approach*, Springer, 2015

[9] Handke J., The Virtual Linguistics Campus, Online: http://linguistics.online.uni-marburg.de/ Accesed 10.25.2018

[10] Boersma P., D. Weenink, University of Amsterdam, Online: http://www.fon.hum.uva.nl/praat/ Accesed 10.25.2018