Sprachaktivitätserkennung basierend auf Deep Neural Networks für Anwendungen in Film und Fernsehen

Niko Moritz[1], Jakob Drefs[1], Hannah Baumgartner[1], Jan Rennies[1]

[1] *Fraunhofer IDMT, Hör-, Sprach- und Audiotechnologie, 26129 Oldenburg, Deutschland, Email: jan.rennies@idmt.fraunhofer.de*

Einleitung

Mangelnde Sprachverständlichkeit von Rundfunkbeiträgen ist ein häufiges Thema, insbesondere bei eher aufwendig produzierten Spielfilmen wie zum Beispiel der etablierten Vorabendserie „Tatort". Traditionell unterliegt die Kontrolle der Sprachverständlichkeit der subjektiven Einschätzung in der Postproduktion. Diese subjektive Einschätzung ist fehleranfällig, insbesondere wenn die Sprachinformationen bereits bekannt sind. Eine automatische Bewertung der Sprachverständlichkeit ist wünschenswert, mit heute verfügbaren Werkzeugen allerdings nicht allgemeingültig realisierbar. Eine wesentliche Voraussetzung für eine automatisierte Bewertung der Sprachverständlichkeit ist die zuverlässige Erkennung von aktiver Sprache im Audiomaterial.

Um Sprachanteile von Nicht-Sprachanteilen eines Signals zu unterscheiden, nutzen Algorithmen zur Sprachaktivitätserkennung (*Voice Activity Detection*, kurz VAD) typischerweise Merkmale wie Energiedifferenz, Periodizität oder spektrale Unterschiede und basieren auf heuristischen Ansätzen oder statistischen Modellen. Dieser Beitrag stellt eine auf *Deep Neural Networks* (DNN) basierende, automatische Erkennung von Sprachaktivität (*Speech Activity Detection*, kurz SAD) und deren Evaluation mit authentischem Fernsehaudiomaterial vor.

Realisierung

Abbildung 1 stellt exemplarisch die Struktur der DNN-basierten Sprachaktivitätserkennung (*Speech Activity Detection*, SAD) dar. Verwendet wird ein vollverbundenes, vorwärtsgerichtetes DNN mit 3 versteckten Ebenen (hidden layers) und 480 Neuronen pro Ebene. Die Aktivierungsfunktion in jeder Ebene basiert auf einer 2er Norm, welche eine Dimensionsreduktion von 480 auf 96 Knotenpunkte bewirkt [1]. Am Eingang des DNN werden Mel-Frequenz-Cepstrum-Koeffizienten (*Mel-Frequency Ceptral Coefficients*, MFCCs) als Merkmalsvektoren verwendet. Am Ausgang des DNN sind zwei Neuronen, die für jeden Eingangsvektor auf Sprache bzw. nicht-Sprache entscheiden. MFCC Merkmale können per se die zeitliche Struktur von Sprache nicht erfassen, da ein MFCC Merkmalsvektor aus relativ kurzen Zeitblöcken (ca. 25 ms) berechnet wird. Für eine zuverlässige Erkennung von Sprachlauten ist die Analyse von Amplitudenmodulationen essentiell [2], wobei Zeitfenster mit einer Länge von ungefähr 200 ms oder größer erforderlich sind [3]. Um einen größeren zeitlichen Kontext mit dem DNN-SAD System zu erfassen, werden die MFCC Merkmalsvektoren über ein Zeitfenster von 470 ms zu einem größeren Merkmalsvektor verbunden, bevor sie in das DNN gefüttert werden. Auf diese Weise werden sowohl zeitliche als auch spektrale Eigenschaften des Audiosignals durch das DNN analysiert, auf dessen Basis die Diskriminierung zwischen Sprache und Nicht-Sprache möglich ist. Das DNN-Training ist vergleichbar mit dem „layer-wise backpropagation Algorithmus" [4], in dem jede DNN-Ebene zunächst zufällig initialisiert wird, um das DNN schichtweise aufzubauen und mittels stochastischem Gradientenverfahren (*stochastic gradient descent*, SGD) zu trainieren [5].

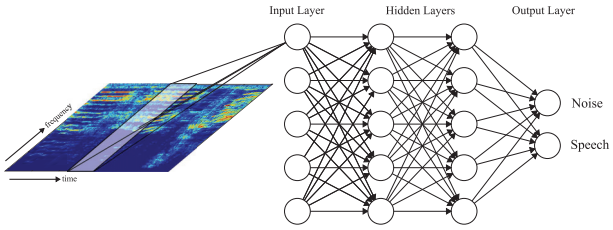

Abbildung 1: DNN-basierte Sprachaktivitätserkennung.

Evaluation

Der vorgestellte DNN-Algorithmus, der aus der Literatur bekannte Algorithmus nach Marzinzik [6] und der ITU–T VAD Standard G.729B [7] werden in unterschiedlichen Testkonditionen miteinander verglichen.

ITU–T G.729B: Der ITU–T VAD Standard G.729B wurde 1996 für Anwendungen rund um Telefonie und Multimedia-Kommunikation entwickelt und wird häufig verwendet, um unterschiedliche VAD-Algorithmen zu vergleichen und zu bewerten. Zur Sprachaktivitätserkennung benutzt der Algorithmus die Energie über alle Frequenzbänder per Zeitfenster, die Energie im unteren Frequenzbereich (bis 1kHz), ein Set von Spektrallinien-Frequenzen (LPC-Koeffizienten) und die Nulldurchgangsraten. Aus den Signalparametern und Energiewerten des Hintergrundsignals werden gleitende Durchschnittswerte berechnet und mit den entsprechenden Werten des aktuellen Frames verglichen. Basierend auf Modellen zu jedem Merkmal wird die Entscheidung, ob es sich um Sprache oder Nicht-Sprache handelt, alle 10 ms gefällt. Der Algorithmus zeigt eine eher geringe Performanz in Konditionen mit Störgeräuschen, insbesondere bei nicht-stationären Störgeräuschen und bei geringen Signal-zuRausch-Abständen (SNR) [8]. Als VAD-Standard kann der ITU-T G.729B aber bei der Erkennung von reinem Sprachmaterial als sehr verlässlich angenommen werden.

Marzinzik: Der von Marzinzik und Kollmeier (2002) vorgeschlagene Algorithmus wurde mit Blick auf

störgeräuschresistente und robuste Merkmale und Entscheidungsregeln entwickelt. Der Ansatz beruht auf spektralen Divergenzmessungen zwischen Sprache und Hintergrund. Der Algorithmus berechnet für jedes Frame die Einhüllende des Leistungsdichtespektrums, und zusätzlich die Leistung des tief- und hochpassgefilterten Signals. Die Differenzen zwischen Maxima und Minima jeder dieser Einhüllenden werden verfolgt und zur Entscheidung für oder gegen eine Sprachpause herangezogen. Hierbei orientiert sich die VAD an Fallunterscheidungen bezüglich der Signaldynamik im hoch- und tiefpassgefilterten Signal. Durch die Anpassung zweier Variablen kann der Algorithmus auf unterschiedliche Signalkategorien angepasst werden bzw. die Falsch-Alarm Rate klein gehalten werden. Die Parameter wurden entsprechend der Beschreibung in [6] gesetzt.

Training der DNN-SAD: Das verwendete Trainingsmaterial umfasst ca. 130h Sprache und 72h Nicht-Sprache. Sprach- und Nicht-Sprach-Segmente in den Trainingsdaten sind in einem Alignment-Prozess mit Hilfe eines Spracherkenners automatisch ermittelt worden. Aus den Alignments wurden die Sprache/Nichtsprache Informationen ermittelt. Zusätzlich wurde der DNN-Erkenner mit Fernsehmaterial trainiert. Hierfür standen in etwa 170 Minuten Audiomaterial aus dem Magazin- und Dokumentarbereich zur Verfügung. Das Material wurde in 10ms-Blöcken unterteilt und innerhalb dieser als Sprache und Nicht-Sprache manuell annotiert. Material, welches zum Training des neuronalen Netzwerkes diente, wurde nicht im Test verwendet.

SNR-Berechnung: Alle Sprachsignale wurden auf einen fixen Sprachpegel gebracht (globale rms-Reduktion, um Clipping zu vermeiden). Bei der Einstellung des Sprachpegels wurde auf die manuellen Annotationen zurückgegriffen, die auch als Referenz zur Berechnung der Hit Rates genutzt wurden. Der Pegel des Störgeräusches wurde über die Abschnitte des Hintergrundsignals berechnet, bei denen Sprache anwesend ist und entsprechend des Ziel-SNR angepasst.

Testmaterial: Als Testmaterial diente ein Werbejingle und die NDR-Fernsehdokumentation „Teilen weltweit“. Die Audiospuren der Dokumentation lagen sowohl getrennt (IT- und Kommentarspur) als auch als Original-Sendemischung vor. Um den Output der VADs zu kontrollieren, wurden „Ground Truth“-Labels manuell erstellt.

Für den Vergleichstest der VADs werden folgende Anwendungsfälle unterschieden:

- Referenzmessungen mit kontrollierten Hintergrundgeräuschen
- Sendemischung mit Vielzahl an Audio-Bestandteilen

Auswertung

Die Güte einer VAD/ SAD wird bewertet bezüglich der *accuracy* (Genauigkeit), mit der Sprache als solche detektiert wird, und Störgeräusch als Störgeräusch klassifiziert wird. Diese Messungen sind als *speech hit rate* und *non-speech hit rate* definiert. Fehlerhafte Klassifizierungen können als unterschiedlich akzeptabel bewertet werden. Kola et al. [9] fanden als aussagekräftigstes Verfahren für die Genauigkeit einer VAD/ SAD die Berechnung der *average hit rate*, einem Mittelwert aus *speech hit rate* und *non-speech hit rate* für einen gegebenen Störgeräuschtypus bei einem gegebenen SNR. Werte nahe bei 1 gelten als Indikatoren für eine hohe Genauigkeit. Durch die gleiche Gewichtung von *speech* und *non-speech hit rate* gibt dieser Fehlerwert allerdings keinen Aufschluss darüber, ob und wann Sprache fälschlicherweise als Störgeräusch klassifiziert wurde.

Referenzmessungen: Für die Referenzmessungen wird die Kommentarspur der NDR Dokumentation anstatt mit der IT-Spur mit technischen Störgeräuschen gemischt. Die Referenzmessungen werden mit weißem Rauschen und dem im Oldenburger Satztest verwendeten modulationsarmen „Olnoise“ [10] durchgeführt. Außerdem wurde noch ein „Cafeteria-Noise“, ein „Applaus-Noise“ und Musik („Swing“: Geige, Kontrabass, Klarinette, Posaune, Schlagzeug mit Gesang) als Hintergrund getestet. Die Algorithmen wurden bei SNR zwischen -15 dB bis +15 dB in Schritten von 5 dB getestet.

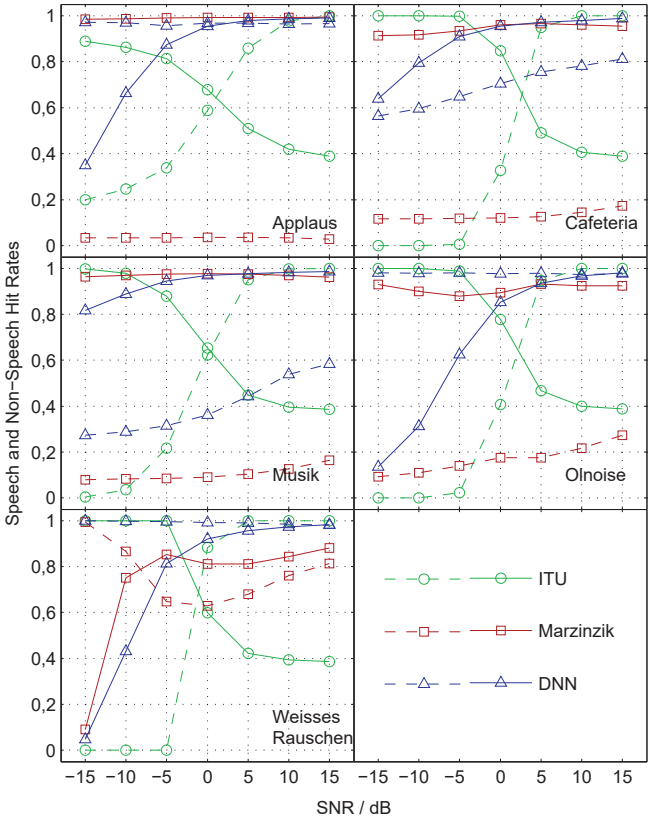

Abbildung 2: Speech Hit Rates (durchgezogene Linien) und Non-Speech Hit Rates (gestrichelte Linien) der DNN-SAD (blau), der Marzinzik-VAD (rot) und des ITU–T G.729B-Algorithmus (grün) bei technischen Hintergrundsignalen in Abhängigkeit vom SNR.

Abbildung 2 zeigt für die Marzinzik-VAD und die DNN-SAD eine klare Verbesserung der Performanz (*accuracy*) mit steigendem SNR, unabhängig vom verwendeten

Störgeräusch. Der Algorithmus nach der ITU–T G.729B zeigt ein eher konträres Verhalten zu den beiden anderen VADs und entscheidet bei schlechten SNR ausschließlich für das Vorhandensein von Sprache. Dadurch erreicht er zwar eine nahezu 100%-ige Speech-Hit Rate, allerdings bei einer äußerst geringen Non-Speech-Hit Rate. Erst ab einem SNR von etwa -5 dB werden tatsächlich „Entscheidungen gefällt“, und nicht mehr nur von Sprache ausgegangen.

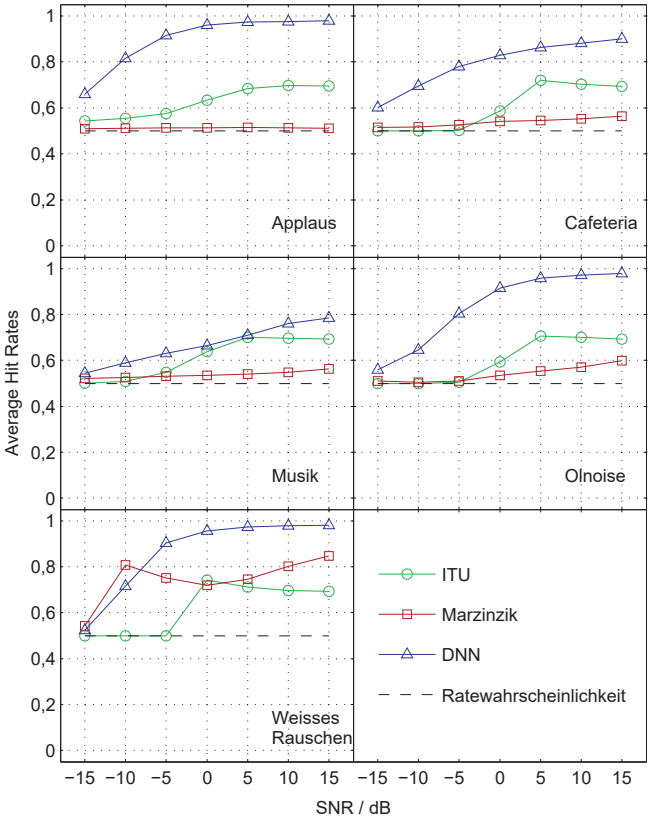

Abbildung 3: Average Hit Rates der DNN-SAD (blau), der Marzinzik-VAD (rot) und des ITU–T G.729B-Algorithmus (grün).

Der Blick auf die Averaged-Hit Rates, also des Mittelwertes von Speech- und Non-Speech-Hit Rate in Abbildung 3 verdeutlicht die tatsächliche Erkennerleistung der Algorithmen. SNRs von kleiner als -5 dB führen in nahezu allen Testkonditionen und alle VADs zu nicht akzeptablen Ergebnissen. Die Marzinzik-VAD und der ITU–T G.729B-Algorithmus liefern Ergebnisse nahe der Ratewahrscheinlichkeit. Die DNN-SAD zeigt in allen Konditionen die besten Ergebnisse, dennoch sind die Resultate für SNRs unter 0dB noch verbesserungswürdig. Ab einem SNR zwischen -5 dB und 0 dB machen sämtliche VADs einen Sprung in ihrer Erkennergüte - außer in der Kondition „Musik“. Hier zeigen sich die Kurven deutlich flacher. Mit Ausnahme der Testkondition Musikßeigt die DNN-SAD bereits ab einem SNR von -5 dB mittlere Trefferquoten von über 80%.

Sendemischung mit Vielzahl an Audio-Bestandteilen

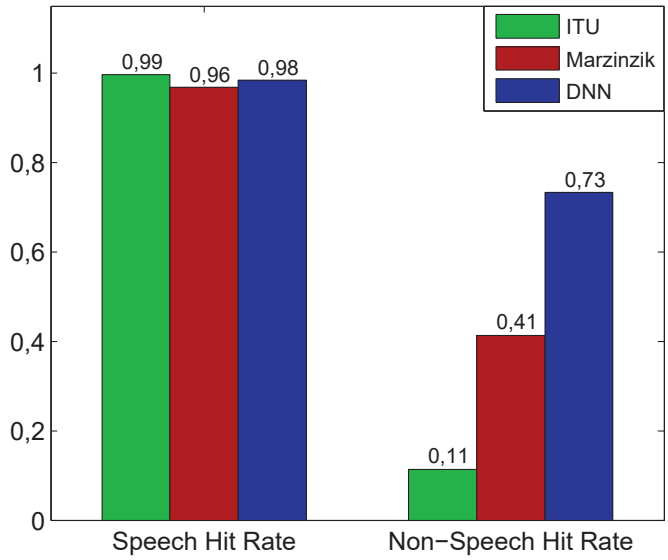

Abbildung 4: Speech / Non-Speech Hit Rates des ITU–T G.729B-Algorithmus (grün), der Marzinzik-VAD (rot) und der DNN-SAD (blau) über die gesamte 25-minütige NDR-Dokumentation „Weltweit teilen“

In einem zweiten Schritt werden ein Werbespot und ein etwa 25-minütiger NDR-Dokumentarfilm analysiert. Werbespot und Dokumentarfilm zeigen die für Film und Fernsehen übliche Variabilität von möglichen Nebengeräuschen (Musik, Effekte, Nachhall, Atmosphäre) bei einem für Fernsehen üblichen Mischungsverhältnis. Sie repräsentiert somit die im Anwendungsfeld spezifische Herausforderung für automatische Sprachaktivitätserkennung.

Abbildung 4 veranschaulicht Speech- und Non-Speech-Hit Raten der drei Algorithmen über den gesamten Dokumentarfilm. Bei der Marzinzik-VAD und dem ITU-Algorithmus verursachen die starken Instationaritäten hohe Fehlalarmraten, beide VADs zeigen zwar eine hohe Speech Hit Rate, allerdings eine geringe Erkennung von Nicht-Sprache Segmenten, entscheiden bei Uneindeutigkeiten also tendenziell zugunsten von Sprache. Die Anzahl der Fehlinterpretationen durch die DNN-SAD fällt deutlich geringer aus, was auch die Ergebnisse für die „average hit rate“ in Tabelle 1 noch einmal verdeutlichen. Die Trefferquote des ITU–T G.729B-Algorithmus bleibt nahe an der Ratewahrscheinlichkeit, die Marzinzik VAD zeigt eine Erkennerrate von 66%. Die DNN-SAD schneidet mit über 85% richtiger Erkennung deutlich besser ab.

Tabelle 1: Average Hit Rates - Fernsehdokumentation

	ITU-VAD	Marzinzik-VAD	DNN-SAD
AHR	0.55	0.66	0.855

Ein zweiter sehr anwendungsnaher Test wurde mit einem etwa 25-sekündigen Werbespot durchgeführt. Abbildung 5 zeigt die Waveform des Jingles. Im unteren Teil der Abbildung sind in unterschiedlichen Farben die

Passagen, die Sprache enthalten (hand-gelabelt), eingetragen und die Passagen des Jingels, in welchen die drei Algorithmen jeweils Sprache ermittelt haben: Der ITU–T G.729B-Algorithmus detektiert nahezu im gesamten Jingle Sprache, eine Unterscheidung zwischen „Speech“ und „Non-Speech“ ist nicht möglich. Die Marzinzik-VAD zeigt keine nennenswerte Verbesserung. Erneut zeigt die DNN-SAD die beste Performanz bei der Diskriminierung von Sprache und Nicht-Sprache.

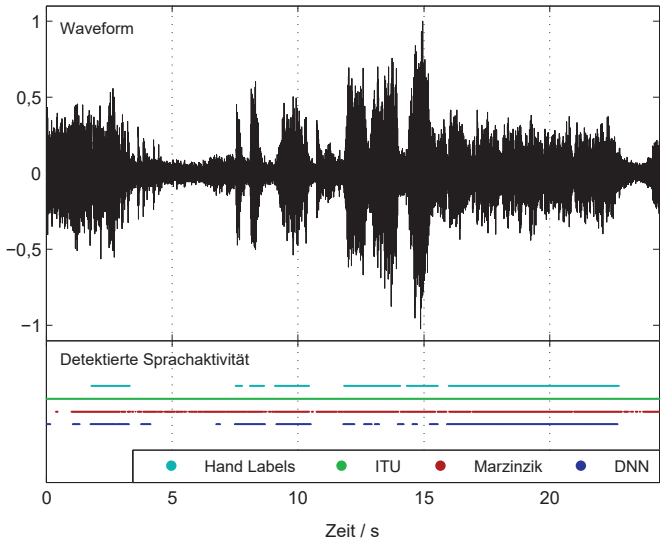

Abbildung 5: Bei Marzinzik- und ITU-Algorithmus verursachen starke Instationaritäten hohe Fehlalarmraten. Geringere Fehlinterpretationen für DNN-SAD.

Zusammenfassung

Die DNN-basierte Sprachaktivitätserkennung zeigt deutlich verbesserte „average hit-rates“ im Vergleich zum Marzinzik- und dem ITU-Algorithmus. Auch eine Optimierung der Parameter im Marzinzik-Algorithmus führte zu keiner nennenswerte Verbesserung der Erkennerleistung (hier nicht gezeigt). Besondere Stärken der DNN-SAD sind bei instationären Störgeräuschen zu erkennen (siehe Testkonditionen: “Applaus“, „Cafeteria“, „Musik“). Während Standardverfahren wie der von Marzinzik hier Ergebnisse nahe der Ratewahrscheinlichkeit aufweisen, kann der DNN-basierte Ansatz weiterhin relativ zuverlässig zwischen Sprache und Nicht-Sprache diskriminieren. Es bleibt zu untersuchen, ob die Leistung des DNN Verfahrens weiter gesteigert werden kann, wenn im Training zum Beispiel auch Musik berücksichtigt wird, was bisher nicht der Fall ist. VADs, deren Schätzung bezüglich unterschiedlicher Störgeräusche und unterschiedlicher SNRs konsistent ist, sind vielseitiger einsetzbar als VADs, deren Erkennergüte stark von Signaleigenschaften oder vom Mischungsverhältnis abhängt. Der Vergleich des neu entwickelten Ansatzes mit etablierten Verfahren zeigt, dass besonders bei instationären Nebengeräuschen eine erhebliche Verbesserung der Erkennungsleistung erreicht wird. Dies kann zukünftig als Vorstufe einer verbesserten Messung der Sprachverständlichkeit eingesetzt werden. Eine Spracherkennung für die Anwendung von Sprachverständlichkeitsmessung oder -verbesserung sollte sehr robust auch mit ungünstigeren SNRs umgehen, da die Ursache für die schlechte Verständlichkeit häufig ein zu gewagtes Mischungsverhältnis ist.

Danksagung

Wir danken dem Norddeutschen Rundfunk für die Bereitstellung von Audiodaten, die im vorliegenden Beitrag für die Evaluierung der VAD Algorithmen verwendet worden sind.

Literatur

[1] X. Zhang, J. Trmal, D. Povey und S. Khudanpur. Improving deep neural network acoustic models using generalized maxout networks, in International Conference on Acoustics, Speech, and Signal Processing, Florence, pp. 215-219, 2014.

[2] T. M. Elliot und F. E. Theunissen. The modulation transfer function for speech intelligibility, PLoS Computational Biology, vol. 5, no. 3, e1000302, 2009.

[3] N. Moritz, J. Anemüller und B. Kollmeier. An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 1926-1937, 2015.

[4] Y. Bengio, P. Lamblin, D. Popovici und H. Larochelle. Greedy layer-wise training of deep networks, in Advances in neural information processing systems, vol. 19, Vancouver, pp. 153-160, 2007.

[5] D. Povey, et al. The Kaldi speech recognition toolkit, in IEEE Automatic Speech Recognition and Understanding Workshop, Hawaii, 2011.

[6] M. Marzinzik und B. Kollmeier. Speech pause detection for noise spectrum estimationby tracking power envelope dynamics, IEEE Trans. Speech Audio Processing, vol. 10, no. 6, pp. 341–351, 2002.

[7] ITU. “Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear –prediction (CS-ACELP). Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70,” International Telecommunication Union, 1996.

[8] H. Farsi, M. Mozaffarian und H. Rahmani. Improving Voice Activity Detection Used in ITU-T G.729.B, in Proc. 3rd WSEAS International Conference on Circuits, Systems, Signal and Telecommunications, Ningbo, China, pp. 11-15, 2009.

[9] J. Kola, C. Espy-Wilson und T. Pruthi. Voice Activity Detection, Merit Bien, 2011.

[10] K. Wagener et al.. Entwicklung und Evaluation eines Satztests in deutscher Sprache I – III: Design, Optimierung und Evaluation des Oldenburger Satztests. Z Audiol 38 (1-3): pp. 4 – 15, 44 – 56, 86 – 95, 1999.