Spectral Subtraction und Hidden Markov Models in der Sprechpausenerkennung

Florian Schleid

30. August 2020

Die Sprechpausenerkennung beziehungsweise Voice Activity Detection dient dazu, für ein eingehendes Signal zu bestimmen, bei welchen Signalabschnitten es sich um Sprache und bei welchen es sich um Hintergrundgeräuschen handelt.
Dadurch ermöglicht die Voice Activity Detection den Rechenaufwand in vielen sprachverarbeitenden Systemen erheblich zu senken, da sprachfreie Abschnitte nicht analysiert werden müssen.
Um die zuverlässige Funktion der sprachverarbeitenden Systeme zu gewährleisten, müssen auch die Voice Activity Detectors eine möglichst hohe Genauigkeit haben. Hierzu werden dann unter anderem Spectral Subtraction und Hidden Markov Models eingesetzt, auf welche im Folgenden eingegangen wird.

Inhaltsverzeichnis

1 Einführung

Voice Activity Detection beziehungsweise Sprechpausenerkennung wird in vielen uns im Alltag begegnenden Geräten eingesetzt. Dabei kann sie als Vorverarbeitungsschritt in Spracherkennern dienen, die beispielsweise in Smart Home Geräten verbaut sind. Sie kann aber auch direkt eingesetzt werden, wie beispielsweise in der Telefontechnik, um die Datenrate zu reduzieren [1, S.1].
Dabei ist die Qualität des Ergebnisses der Voice Activity Detetor auch stark davon abhängig, wie hoch die Qualität des Eingangssignals ist. Ist das Signal sehr verrauscht, ist es für den VAD auch schwer die Sprache herauszufiltern. Die Qualität des Signals wird dabei durch die Signal-to-Noise ratio (SNR) angegeben. Eine niedrigere SNR, bedeutet, dass viel Rauschen im Signal vorhanden ist. Um trotzdem zufriedenstellende Ergebnisse vom VAD gewinnen zu können, befassen wir uns mit Spectral Subtraction und Hidden Markov Modellen als zusätzlichem Schritt in der Verarbeitung des Eingangssignals, um die Genauigkeit zu verbessern.

2 Allgemeine Funktionsweise eines Voice Activity Detector's

Im Folgenden wird erläutert, wie ein Voice Activity Detector allgemein funktioniert, um eine Grundlage für die Erläuterung von Hidden Markov Modellen und Spectral Subtraction im Kontext der Sprechpausenerkennung zu schaffen.
Zunächst muss das VAD-System trainiert werden, wobei die internen Parameter des Systems angepasst werden. Dies kann beispielsweise der Schwellwert für die Entscheidung, ob Sprache vorliegt sein oder statistische Parameter wie die Varianz und der Mittelwert oder auch die durchschnittliche Zero Crossing Rate. Dazu benötigt das System Signalvektoren, die zu Hintergrundrauschen gehören. Diese sollten keine Sprache enthalten [2, S.1].
Nun können die Signalabschnitte, die untersucht werden sollen, in das System eingegeben werden. Auf Grundlage der eingestellten Parameter wird entschieden, ob es sich bei einem Abschnitt um Sprache handelt oder um Hintergrundgeräusche.
Wenn es sich um einen Abschnitt ohne Sprache handelt, wird dieser wieder zum Verbessern der Entscheidungsparameter eingesetzt [2, S.1].

3 Funktionsweise von Spectral Subtraction

In diesem Abschnitt soll nun darauf eingegangen werden, wie Spectral Subtraction funktioniert.
Man kann annehmen, dass sich das Eingabesignal $y(m)$ in den Teil der Sprache $x(m)$ und in den Teil der Störgeräusche $n(m)$ aufteilt. Es ergibt sich

$$y(m) = x(m) + n(m)$$

für das von Rauschen durchsetzte Sprachsignal [2, S.3].
Die spektrale Leistungsdichte eines Signals beschreibt den Zusammenhang der Frequenz eines Signals, bezogen auf dessen Leistung [3, S.64]. Indem man den Durchschnitt der spektralen Leistungsdichte des Noise Signals von der spektralen Leistungsdichte des Eingabesignals abzieht, erhält man die spektrale Leistungsdichte des Signals, in dem der Anteil an Rauschen reduziert ist. Dabei wird zudem ein Faktor α eingesetzt, der die Stärke der Subtraktion beschreibt. Es ergibt sich folgende Gleichung [2, S.3]:

$$|\hat{X}(f)|^2 = |Y(f)|^2 - \alpha|N(f)|^2$$

Um den Durchschnitt der spektralen Leistungsdichte des Noise Signals $|N(f)|^2$ zu ermitteln, nimmt man eine Anzahl K an bekannten Noise-Abschnitten aus den Trainingsdaten und mittelt deren spektrale Leistungsdichte:

$$|N(f)|^2 = \frac{1}{K}\sum_{i=0}^{K-1}|N_i(f)|^2$$

Dabei beschreibt $|N_i(f)|^2$ das spektrale Leistungsdichtespektrum des i-ten Noise-Abschnittes.
Um nun aus der spektralen Leistungsdichte $|\hat{X}(f)|$, welche ja die Leistung in Bezug zur Frequenz setzt (frequency domain), wieder zurück zur Beschreibung der Leistung des Signals im Bezug zur Zeit (time domain) zu erhalten, wendet man die inverse Fourier Transformation an [4, S.2], wobei man ebenfalls die Phase der Noise-Abschnitte berücksichtigen muss. Es ergibt sich:

$$\hat{x}(m) = \sum_{i=0}^{K-1}|\hat{X}(k)|e^{j\theta_y k}e^{-j\frac{2\pi}{N}km}$$

Dabei ist θ_y die Phase des Noise-Signals der Frequenz $Y(k)$ [2, S.3].
Das entrauschte Signal $\hat{x}(m)$ kann nun als Eingabesignal einem VAD übergeben werden, der nun aufgrund der höheren SNR bessere Ergebnisse erzielen kann, wie wir später bei der Evaluation sehen werden.

4 Ein Voice Activity Detector basierend auf einem Hidden-Markov Model

In diesem Abschnitt wird ein Sprechpausenerkenner vorgestellt, der auf einem Hidden-Markov Model basiert.
Um eine Entscheidung zu erzielen, ob ein Abschnitt eines Signals Sprache enthält oder nicht, muss der VAD den Abschnitt bewerten. Wenn diese Bewertung eine bestimmte Schwelle überschreitet, wird der Abschnitt als Sprache betitelt.
Die Bewertung eines Abschnittes soll durch die Funktion $V^{soft}(t)$ vorgenommen werden. Diese sei gegeben durch [5, S.4]:

$$V^{soft}(t) = p(H_1|y_t) = \frac{\kappa_t L(y_t)}{1 + \kappa_t L(y_t)}$$

Dabei ist y_t der Signalabschnitt der aktuell betrachtet werden soll [5, S.2], H_1 ist die Hypothese, dass in dem Abschnitt Sprache vorhanden ist [5, S.3], κ_t sei das Verhältnis des bisherigen Anteils von Abschnitten mit Sprache im Signal, zum bisherigen Anteil von sprachfreien Abschnitten im Signal und $L(y_t)$ ist die Likelihood ratio zum Zeitpunkt t [5, S.4].
Je größer der Anteil der Abschnitte mit Sprache im bisherigen Signal ist, und je größer die Wahrscheinlichkeit ist, dass im aktuellen Frame Sprache vorhanden ist, welche durch die Likelihood ratio $L(y_t)$ angegeben wird [5, S.4], desto kleiner ist der Einfluss der 1 im Nenner der Gleichung auf das Ergebnis des Bruches, welches sich daher 1 annähert. Da $\kappa_t > 0$ und $L(y_t) > 0$ gilt, kann Ergebnis auch nicht negativ werden und es gilt folglich $V^{soft}(t) \in [0, 1]$
Um κ_t zu bestimmen benötigen wir die a priori Wahrscheinlichkeit $p_t(H_1)$, die die Wahrscheinlichkeit, dass in einem Abschnitt Sprache vorliegen könnte, auf Basis der bis zum Zeitpunkt t erhaltenen Ergebnisse angibt. Wenn wir $p_t(H_1)$ kennen, können wir mit $p_t(H_0) = 1 - p_t(H_1)$ die Gegenwahrscheinlichkeit berechnen, wobei H_0 die Hypothese ist, dass im betrachteten Abschnitt keine Sprache vorliegt [5, S.4].

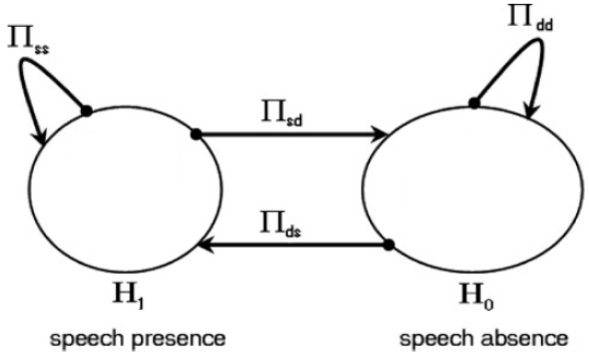
Für $p_t(H_1)$ gilt die Gleichung [5, S.4]

$$p_t(H_1) = \Pi_{ds} p(H_0|y_{t-1}) + \Pi_{ss} p(H_1|y_{t-1})$$

Π_{ds} gibt die Wahrscheinlichkeit dafür an, dass von einem Abschnitt ohne Sprache ein Übergang in einen Abschnitt mit Sprache stattfindet. Π_{ss} dementsprechend die Wahrscheinlichkeit, dass ein Abschnitt Sprache enthält, wenn der Vorherige ebenfalls Sprache enthalten hat. $p(H_0|y_{t-1})$ und $p(H_1|y_{t-1})$ ergeben sich aus den Ergebnissen von V^{soft} für den vorherigen Abschnitt. Π_{ds} und Π_{ss} ergeben sich aus den Trainingsdaten [5, S.4]. Nun da wir κ_t berechnen können benötigen wir noch die Likelihood ratio $L(y_t)$, um $V^{soft}(t)$ berechnen zu können.
$L(y_t)$ ist das Verhältnis der Wahrscheinlichkeit, dass das gemessene Signal y_t vorliegt, wenn es sich um Sprache handelt, zu der Wahrscheinlichkeit, dass das Signal y_t vorliegt, wenn keine Sprache vorhanden ist [5, S.4]:

$$L(y_t) = \frac{P(y_t|H_1)}{P(y_t|H_0)}$$

Nun kann $V^{soft}(t)$ berechnet werden.
Diese Berechnungen entsprechen einem Hidden-Markov Modell mit zwei Zuständen. Der erste Zustand ist H_1, also dass Sprache vorhanden ist, und der zweite Zustand ist H_0, also dass keine Sprache vorhanden ist. Die Übergänge zwischen diesen Zuständen beschreiben die Übergangswahrscheinlichkeiten Π_{sd} und Π_{ds}. Die Übergänge von einem Zustand zurück auf sich selbst, sind entsprechend Π_{ss} und Π_{dd}[5, S.4][6, S.4].
Die Verwendung eines HMM führt dazu, dass Sprache mit geringer Lautstärke seltener als Rauschen klassifiziert wird, da die Entscheidung nicht nur aufgrund der aktuellen Ergebnisse, sondern auch aufgrund des vorherigen Abschnittes getroffen wird [6, S.4].



Auf der Grundlage von $V^{soft}(t)$ und der Schwelle T^{VAD} kann nun die endgültige Entscheidung des VAD $V^{hard}(t) \in \{0, 1\}$ bestimmt werden[5, S.4]:

$$V^{hard}(t) = \begin{cases} 1 \text{ wenn} V^{soft}(t) \geq T^{VAD} \\ 0 \text{ sonst} \end{cases}$$

5 Bewertung der Methoden

Nachdem die beiden Methoden nun vorgestellt wurden, soll im Folgenden deren Wirksamkeit dargestellt werden.
Zunächst soll der Einsatz von Hidden-Markow Modellen in VADs evaluiert werden. Dazu wird der VAD mit zwei anderen Detektoren bei verschiedenen SNR verglichen. Diese sind zum einen der ITU-T G.729B standard Voice Activity Detektor [7] und ein Voice Activity Detektor basierend auf einem Laplace-Gauß'schem statistischen Modell [6].

5.1 Bewertung von HMMs in VADs

Der ITU-T G.729B standard VAD wird im Folgenden mit G.729 abgekürzt und der zweite VAD mit LapGa.
Die Signale, die zur Bewertung von den VADs verarbeitet werden, enthalten vier verschiedene Arten von Hintergrundgeräuschen: 1. Bürogeräusche, 2. Gemurmel, 3. Cockpitgeräusche eines F16 Jets und 4. Maschinengewehrfeuer. Dabei liegen Signal to Noise Ratio von 0, 5, 10 und 15 dB vor [5, S.5]. Für den Zustand des Hidden-Markov Modell, in dem Sprache vorliegt, kann man als Modelle entweder λ_s^{mfc}, λ_y^{mfc} oder λ_y^{mfs} heranziehen. Diese beinhalten die Parameter für das Hidden-Markov Modell. λ_s^{mfc} ist ein Modell, dass auf sauberer Sprache in der Darstellungsform im MFC (Mel-Frequency Cepstral) -Bereich basiert. Bei λ_y^{mfc} handelt es sich um Sprache mit Hintergrundgeräuschen und bei λ_y^{mfs} ebenfalls um Sprache mit Hintergrundgeräuschen, aber im MFS (Mel-Frequency Spectral) -Bereich [5, S.2, S.6].
Zum Trainieren des Systems werden 300 verschiedene Sätze von verschiedenen Sprechern verwendet. Zum anschließenden Testen der Genauigkeit liegen 10 Sätze vor. Zum

Bewerten der Performance der VADs wird die Spracherkennungsrate

$$R_{det} = \frac{\text{Anzahl der korrekt als Sprache klassifizierten Abschnitte}}{\text{Anzahl der tatsächlich Sprache enthaltenden Abschnitte}}$$

und die Fehlalarmrate

$$R_{fals} = \frac{\text{Anzahl der fälschlicherweise als Sprache klassifizierten Abschnitte}}{\text{Gesamtanzahl der Abschnitte, die keine Sprache enthalten}}$$

verwendet. Der Schwellwert des VAD T^{VAD} ist auf 0,1 gesetzt [5, S.5].
Wie man Tabelle 1 auf der nächsten Seite entnehmen kann, ergeben sich bei λ_s^{mfc} niedrigere R_{det} und R_{fals} als bei λ_y^{mfc} und λ_y^{mfs}. Dies liegt daran, dass das verrauschte Signal den Modellen, in denen Hintergrundgeräusche vorhanden sind, grundsätzlich stärker ähnelt, als dem Modell mit ausschließlich sauberer Sprache. Dadurch ist auch die Likelyhood ratio $L(y_t)$ bei λ_y^{mfc} und λ_y^{mfs} größer, was wiederum zu einem größeren Wert für $V^{soft}(t)$ führt [5, S.6].
Bei der Verwendung von λ_s^{mfc} ergibt sich allerdings das Problem, dass R_{det} bei niedriger SNR stark abnimmt (siehe Tabelle 1 auf der nächsten Seite). Dies ist bei λ_y^{mfc} und λ_y^{mfs} nicht der Fall. Sie erreichen eine konstant hohe R_{det}, weshalb auch der Durchschnitt der Spracherkennungsraten bei 96,7% und 98,6% liegt, während der von λ_s^{mfc} nur bei 65,4% liegt (Tabelle 1 auf der nächsten Seite).
Gegenüber den beiden zum Vergleich herangezogenen VADs erreichen λ_y^{mfc} und λ_y^{mfs} vor allem bei niedrigen SNR Leveln bei den Rauschtypen Büro und F16-Jet deutlich höhere Spracherkennungsraten R_{det}. Vor allem der LapGa VAD erreicht nur noch geringe R_{det} bei diesen Geräuschtypen (Tabelle 1 auf der nächsten Seite). Auch beim Gemurmel und beim Maschinengewehrfeuer erreichen sie ähnliche oder höhere Ergebnisse bezüglich R_{det} (Tabelle 1 auf der nächsten Seite).
Das Problelm beim LapGa, G729 aber auch bei λ_s^{mfc} ist, dass sie in stark verrauschten Signalen Probleme mit schnellen Veränderungen im Hintergrundrauschen haben [5, S.6].
Man kann allerdings auch eine höhere Fehlalarmrate R_{fals} bei λ_y^{mfc} und λ_y^{mfs} feststellen. Dies schlägt sich etwa in den Durchschnittswerten nieder, die mit 65,4% und 70,9% deutlich höher liegen als die vom LapGa und vom G729 mit 56,7% und 60.4% (Tabelle 1 auf der nächsten Seite). Das Verhältnis von R_{det} und R_{fals} wird von dem Wert für T^{VAD} bestimmt. Ein höherer Wert führt zu weniger falschen Klassifizierun-

gen als Sprache, aber auch dazu, dass mehr undeutliche Sprachsignale als Rauschen interpretiert werden. Gleichermaßen resultiert eine niedrigere Schwelle in höheren R_{det} und R_{fals} [5, S.6].

Tabelle 1: Ergebnisse der VADs für verschiedene Arten von Rauschen und bei verschiedenen SNR in Prozent [5, S.7]

Rauschen	SNR, dB	λ_s^{mfc}		λ_y^{mfc}		λ_y^{mfs}		G.729		LapGa	
		R_{fals}	R_{det}	R_{fals}	R_{det}	R_{fals}	R_{det}	R_{fals}	R_{det}	R_{fals}	R_{det}
Büro	0	6	16	61	93	79	100	46	73	19	27
	5	6	50	58	97	72	99	42	83	19	34
	10	9	69	60	96	68	99	48	87	41	55
	15	12	80	62	98	58	99	48	89	65	93
Gemurmel	0	1	25	70	95	73	97	71	90	65	90
	5	1	57	65	97	66	98	85	94	65	94
	10	4	79	51	98	61	99	83	95	65	99
	15	13	86	56	97	66	98	62	94	65	98
Maschinengewehr	0	23	72	50	96	88	100	83	93	85	99
	5	29	81	65	97	81	99	82	93	76	99
	10	44	90	76	98	88	100	80	92	70	99
	15	57	95	84	98	88	100	77	92	68	99
F16	0	9	31	71	93	70	99	37	75	29	35
	5	10	56	72	98	64	97	42	84	44	54
	10	11	75	75	98	59	97	38	86	65	92
	15	13	82	71	98	54	96	42	90	65	98
Durchschnitt		15,5	65,3	65,4	96,7	70,9	98,6	60,4	88,1	56,7	79,3

5.2 Bewertung von Spectral Subtraction

Zur Bewertung von Spectral Subtraction als Vorverarbeitungsschritt für einen Voice Activity Detector vergleichen wir die Ergebnisse eines VAD mit und ohne vorgeschalteter Spectral Subtraction. Dabei werden drei verschiedene Arten von Rauschen verwendet: 1. Weißes Rauschen, 2. Pinkes Rauschen und 3. Braunes Rauschen [2, S.4, S.5]. Bei weißem Rauschen ist die Rauschleistung in einem Abschnitt unabhängig von der Frequenz [8, S.38]. Bei pinkem Rauschen nimmt die Rauschleistung mit steigender Frequenz ab, sodass ein Mensch alle Frequenzbereiche als etwa gleich laut empfindet

[9]. Braunes Rauschen verhält sich ähnlich wie pinkes Rauschen, die Rauschleistung sinkt jedoch quadratisch bei ansteigender Frequenz [10].

Tabelle 2: Ergebnisse für weißes Rauschen [2, S.5]

Eingabesignal mit weißem Rauschen	
Ergebnis des VAD ohne Spectral Subtraction	
Eingabesignal nach Spectral Subtraction	
Ergebnis des VAD mit Spectral Subtraction	

Wie man in Tabelle 2 sehen kann, liegen beim Ergebnis des VAD ohne vorherige Spectral Subtraction einige Bereiche, beispielsweise zu Beginn und zum Ende des Signals vor, die tatsächlich Rauschen sind und vom VAD als Sprache erkannt wurden. Diese sind im Ergebnis des VAD mit Spectral Subtraction nicht mehr vorhanden.

Tabelle 3: Ergebnisse für pinkes Rauschen [2, S.5]

Eingabesignal mit pinkem Rauschen	
Ergebnis des VAD ohne Spectral Subtraction	
Eingabesignal nach Spectral Subtraction	
Ergebnis des VAD mit Spectral Subtraction	

Man kann auch bei pinkem Rauschen (Tabelle 3) erkennen, dass Spectral Subtraction den Effekt, dass starkes Rauschen als Sprache klassifiziert wird, abschwächt. Bereits nachdem Spectral Subtraction auf das Eingabesignal angewendet wurde, lässt sich erkennen, dass das Signal deutlich entrauscht wurde, was zu einer höheren SNR und

damit auch zu einem besseren VAD-Ergebnis führt.

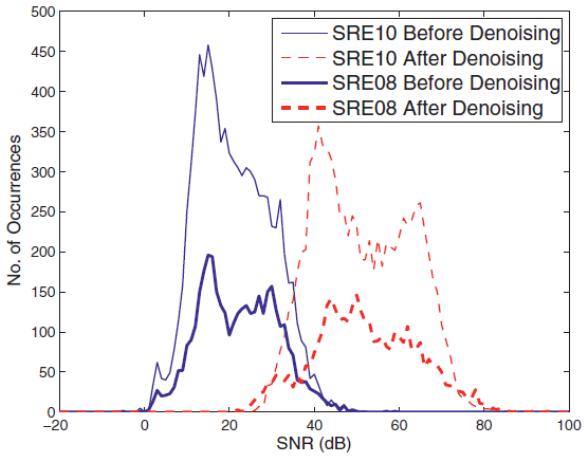
Tabelle 4: Ergebnisse für braunes Rauschen [2, S.5]

Eingabesignal mit braunem Rauschen	
Ergebnis des VAD ohne Spectral Subtraction	
Eingabesignal nach Spectral Subtraction	
Ergebnis des VAD mit Spectral Subtraction	

Auch bei braunem Rauschen (Tabelle 4) kann man erkennen, dass im Ergebnis weniger Störungen sind.
Es können also für alle drei Arten von Rauschen Verbesserungen im Ergebnis festgestellt werden.

Insgesamt kann man sagen, dass der Einsatz von Spectral Subtraction die SNR in den jeweiligen Signalen erhöht, wie sich auch in Abbildung 1 erkennen lässt.

Abbildung 1: Histogramm der SNR von den NIST speaker recognition evaluations (SRE) Datensätzen aus 2008 und 2010 vor und nach dem Anwenden von Spectral Subtraction [11, S.5]



6 Fazit

Wir haben gesehen, dass sich durch den Einsatz von Hidden-Markow Modellen hohe Erkennungsraten erzielen lassen. Gleichzeitig waren aber auch die Fehlalarmraten höher als die der referenzierten Voice Activity Detektoren. Auch wenn man diese beiden Größen durch eine Veränderung des Schwellwerts T^{VAD} abändern kann, spielt hier auch die Zielanwendung eine bedeutende Rolle. Wenn man den VAD beispielsweise nur zur Verbesserung der Signalqualität verwenden möchte, spielt eine hohe Fehlalarmrate eher eine untergeordnete Rolle, während es sehr bedeutend ist, dass möglichst alle Sprache als solche erkannt wird.
Man konnte anhand der Ergebnisse auch erkennen, dass eine hohe SNR vorteilhaft für das Ergebnis des VADs ist. Diese wird durch den Einsatz von Spectral Subtraction erhöht, weshalb es als sinnvoll erscheint Spectral Subtraction als Vorverarbeitungsschritt einzusetzen.

Literatur

[1] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano. Performance evaluation and comparison of g.720/amr/fuzzy voice activity detectors. 2002.

[2] Tejus Adiga M and Rekha Bhandarkar. Improving single frequency filtering based voice activity detection (vad) using spectral subtraction based noise cancellation. 2016.

[3] Winfrid G. Schneeweiss. *Korrelationsfunktion und spektrale Leistungsdichten in linearen Systemen.* Springer, 1974.

[4] Kim Mey Chew, Rubita Sudirman, Nasrul Humaimi Mahmood, Norhudah Seman, and Ching Yee Yong. Human brain microwave imaging signal processing: Frequency domain (s-parameters) to time domain conversion. 2013.

[5] H Veisi and H Sameti. Hidden-markov-model-based voice activity detector with high speech detection rate for speech enhancement. April 2011.

[6] Saeed Gazor and Wei Zhang. A soft voice activity detector based on a laplacian-gaussian model. September 2003.

[7] Adil Benyassine, Eyal Shlomot, Huan-Yu Su, Dominique Massaloux, Claude Lamblin, and Jean-Pierre Petit. Itu-t recommendation g.729 annex b: A silence compression scheme for use with g.729 optimized for v.70 digital simultaneous voice and data applications. 1997.

[8] R. Müller. *Rauschen.* Springer-Verlag, 1979.

[9] 1/f-rauschen. https://de.wikipedia.org/wiki/1/f-Rauschen.

[10] $1/f^2$-rauschen. https://de.wikipedia.org/wiki/1/f^2-Rauschen.

[11] Man-Wai Mak and Hon-Bill Yu. A study of voice activity detection techniques for nist speaker recognition evaluations. 2013.