Schätzung des Signal-Rauschabstandes durch Analyse von Amplitudenmodulationen

Jürgen Tchorz und Birger Kollmeier
AG Medizinische Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg, 26111 Oldenburg
tch@medi.physik.uni-oldenburg.de

I. Einleitung

Die automatische Klassifikation akustischer Signale ist eine wichtige Voraussetzung für eine Reihe technischer Anwendungen. Eine Störgeräuschunterdrückung für automatische Spracherkennung beispielsweise erfordert eine verlässliche Detektion von Sprachpausen bzw. des Störgeräuschanteils in verschiedenen Frequenzbändern. Üblicherweise wird dabei das Störgeräuschspektrum in detektierten Sprachpausen ermittelt. Für eine wirksame Störgeräuschunterdrückung müssen dabei zwei Voraussetzungen erfüllt sein: 1.) möglichst alle Sprachpausen werden fehlerfrei detektiert, und 2.) während Sprachaktivität ist das Störgeräusch stationär, damit die Schätzung nicht fehlerhaft ist. In der Realität sind beide Voraussetzungen oftmals nicht hinreichend erfüllt, so daß technische Störgeräuschunterdrückungsverfahren in vielen Situationen nur unbefriedigende Resultate liefern.

Das in diesem Beitrag vorgestellte Verfahren ermöglicht eine direkte Schätzung des SNR in verschiedenen Frequenzbändern in kurzen Analysefenstern, ohne dabei eine zeitweise Stationarität des Störgeräusches oder die Abwesenheit von Sprache vorauszusetzen. Dabei werden aus dem Eingangssignal spektrale und zeitliche Merkmale extrahiert, welche neurophysiologisch motiviert sind. In den entstehenden komplexen Mustern bilden sich Sprache und Störgeräusche in einer charakteristischen Art und Weise ab. Durch eine Mustererkennung wird eine Schätzung des aktuellen SNR in einzelnen Frequenzbändern ermöglicht.

II. Signalverarbeitung & Mustererkennung

Zur Schätzung des SNR werden aus dem Eingangssignal sogenannte Amplitudenmodulationsspektrogramme generiert (AMS). Diese Muster sind motiviert durch neurophysiologische Versuche zur Verarbeitung von Amplitudenmodulationen in höheren Stufen der Gehörbahn. Langner und Schreiner [1, 2] wiesen eine topologische Anordnung von Neuronen in Abhängigkeit ihrer Best-Modulationsfrequenz nach. Diese topologische Anordung steht nahezu senkrecht zur bekannten tonotopen Anordnung bezüglich der Best-Mittenfrequenzen der Neuronen. Zur technischen Nachbildung solcher zweidimensionalen, frequenz-zeitlichen Muster [3] wird das mit 16 kHz abgetastete Eingangssignal zunächst in seinem Langzeitpegel ausgeglichen. Das Spektrum in 4 ms Fenstern (Vorschub: 0.25 ms) wird durch FFT berechnet und das Einhüllendensignal in jedem Frequenzband durch Quadrierung der Amplitude gebildet. Anschließend wird in jedem Band über Fenster von 32 ms Länge mittels FFT das Modulationsspektrum berechnet. Durch Summation entsprechender Bänder werden beide Achsen logarithmisch umskaliert. Die Mittenfrequenzen reichen dabei von 100-7800 Hz, das Modulationsspektrum von 50-400 Hz, in jeweils 15 Bändern. Abschließend werden die Amplituden logarithmisch komprimiert. Jedes AMS Muster repräsentiert spektrale und zeitliche Eigenschaften des Signals in einem 32 ms Analysefenster. In Abb. 1 (links) ist ein AMS Muster dargestellt, das aus stimmhafter Sprache erzeugt wurde. Deutlich zu sehen ist die verstärkte Energie bei einer Modulationsfrequenz im Bereich der Grundfrequenz (ca. 110 Hz), sowie bei den ersten beiden Harmonischen. Das rechte AMS Muster wurde aus sprachsimulierendem Rauschen erzeugt. Zu erkennen ist das typische Abfallen der Energien zu höheren Frequenzen hin, aber keinerlei Struktur in den Modulationsspektren.

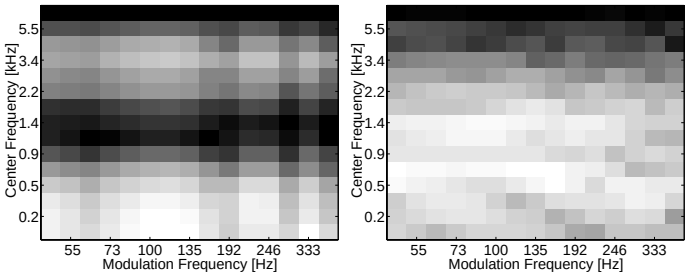

Abb. 1: *Beispiele für AMS Muster*

Die charakteristischen Unterschiede zwischen Sprache und Störgeräuschen in AMS Mustern werden durch ein Mehrschichtperzeptron ausgewertet. Es besteht aus einer Eingangsschicht mit 225 Neuronen (15 * 15, die Auflösung von AMS Mustern), einer verdeckten Schicht mit 40 Neuronen, und einer Ausgabeschicht mit 15 Neuronen. Jedes Ausgabeneuron repräsentiert dabei den SNR in einem Frequenzband. In der Trainingsphase wird der tatsächliche SNR des Eingangssignals für jedes Analysefenster, welches anschließend in ein AMS Muster umgewandelt wird, in allen 15 Bändern gemessen (Sprache und Störgeräusch liegen getrennt vor). Die gemessenen SNRs dienen als Zielaktivität der jeweiligen Ausgangsneuronen. In der Testphase werden "unbekannte" Eingangssignale in AMS Muster umgewandelt. Die auftretenden Aktivitäten der Ausgabeneuronen dienen nun als Schätzung des aktuellen SNR in den jeweiligen Bändern.

III. Experimente

A. Material

Zum Trainieren des Neuronalen Netzwerkes wurde eine insgesamt 72-minütige Mischung aus Sprache und Störgeräusch verwendet. Der Langzeit-SNR betrug dabei 2.5 dB, aber der *lokale* SNR in den 32 ms Analysefenstern unterlag starken Schwankungen (z.B. in Sprachpausen). Das Trainings-Sprachmaterial wurde dem Phondat-Korpus entnommen [4] und bestand aus 2110 Sätzen, gesprochen von 190 männlichen und 210 weiblichen SprecherInnen. Als Störgeräusch wurden 41 verschiedene natürliche Geräuschquellen verwendet. Als Testmaterial diente eine 36-minütige Mischung aus Sprache (Phondat, je 100 SprecherInnen) und 54 natürlichen Geräuschquellen, die jeweils nicht im Trainingsmaterial enthalten waren. Das Perzeptron wurde mit 100 Iterationen trainiert.

B. Ergebnisse

Eine beispielhafte schmalbandige SNR Schätzung mit dem vorgestellten Verfahren ist in Abb. 2 gezeigt. Beim Eingangssignal handelte es sich um eine Mischung aus Sprache und Bohrmaschinenlärm. Dargestellt sind der tatsächliche SNR (durchgezogene Linien) und der geschätzte SNR (gestrichelt) in 7 von 15 Frequenzbändern. In hohen Frequenzbändern (oben) ist der SNR wegen des hochfrequenten Störgeräusches relativ schlecht. Insgesamt wird eine recht gute Übereinstimmung zwischen gemessenem und geschätztem SNR erreicht, abgesehen von den oberen Bändern, in denen es zu deutlichen Abweichungen kommt.

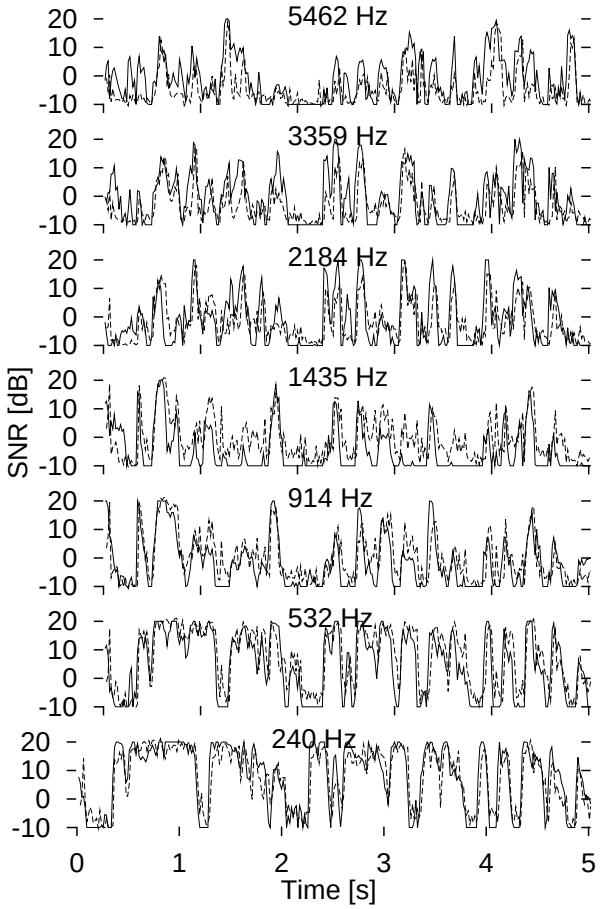

Abb. 2: *Beispiel für SNR Schätzung*

Eine quantitative Beurteilung der SNR Schätzung wird erreicht durch Berechnung der mittleren Abweichung D zwischen dem gemessenen SNR a_i und dem geschätzten SNR e_i über N AMS Muster (mit Index i):

$$D = \frac{1}{N} \sum_{i=1}^{N} |a_i - e_i| \qquad (1)$$

Die mittlerer Abweichung D wurde über alle AMS Muster des oben beschriebenen Test- und Trainingsmaterials in allen 15 Frequenzbändern berechnet und in Abb. 3 dargestellt. Für die Testdaten (durchgezogene Linie) zeigt sich, daß die SNR Schätzung zu höheren Frequenzbändern hin ungenauer wird. In den meisten Bändern jedoch ist die Genauigkeit der Schätzung nur geringfügig schlechter als für das Trainingsmaterial (gestrichelte Linie), das Perzeptron generalisiert also relativ gut. Die mittlere Abweichung zwischen gemessenem und geschätztem SNR über alle Frequenzbänder liegt bei 5.4 dB.

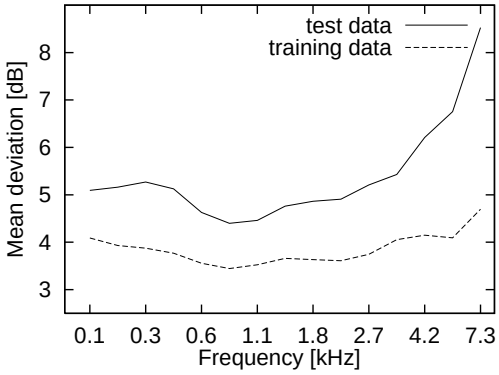

Abb. 3: *Mittlere Abweichung zwischem gemessenem und geschätztem SNR.*

IV. Diskussion

Der vorgestellte Algorithmus erlaubt eine Schätzung des SNR in 15 verschiedenen Frequenzbändern in voneinander unabhängigen Analysefenstern von 32 ms Länge. Dabei wird der SNR direkt geschätzt, und keinerlei Annahmen über die Stationarität des Störgeräusches sind erforderlich. Die dazu notwendigen Informationen sind in den beschriebenen AMS Mustern enthalten, welche sowohl spektrale, als auch zeitliche Charakteristika des Eingangssignals beinhalten. Ähnliche Muster wurden bereits zur binauralen Störgeräuschunterdrückung [3] und Vokaltrennung [5] eingesetzt. Durch die Schätzung des SNR in verschiedenen Bändern der Analysefenster sind die Voraussetzungen geschaffen für eine monaurale Störgeräuschunterdrückung durch Abschwächung von Bändern mit schlechtem SNR. Im Gegensatz zu herkömmlichen Verfahren [6, 7] ist die Anzahl der zur Verfügung stehenden Bänder jedoch sehr viel geringer, was die Unterdrückung von Störgeräuschen mit sehr "ungleichmäßigem" Spektrum erschwert. Andererseits wird dadurch die Entstehung einiger unerwünschter Artefakte verhindert, wie informelle Tests zeigten.

Teile dieser Arbeit wurden gefördert von der Europäischen Union (TIDE/SPACE)

Literatur

1. G. Langner and C.E. Schreiner, "Periodicity coding in the inferior colliculus of the cat. i. neuronal mechanisms.," *J. Neurophysiol.*, vol. 60, pp. 1799–1822, 1988.
2. G. Langner, M. Sams, P. Heil, and H. Schulze, "Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: evidence from magnetoencephalography," *J. Comp. Physiol. A*, vol. 181, pp. 665–676, 1997.
3. B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1593–1602, 1994.
4. K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, and W. Thon, "Handbuch zur datenaufnahme und transliteration in tp14 von verbmobil-3.0.," Tech. Rep., Verbmobil-Technischer Report, 1994.
5. D. Yang, G. F. Meyer, and W. A. Ainsworth, "A neural model for auditory scene analysis," *J. Acoust. Soc. Am.*, vol. 105, no. 2, pp. 1092, 1999.
6. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.
7. Y. Ephraim and M. Malah, "Speech enhancement using a minimum mean-sqare error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.