# MEL-FREQUENZANPASSUNG DER MINIMUM VARIANZ DISTORTIONLESS RESPONSE EINHÜLLENDEN

*Matthias Wölfel*

*Institut für Logik, Komplexität und Deduktionssysteme*
*Universität Karlsruhe (TH)*

*wolfel@ira.uka.de*

**Abstract:** Bilineare Abbildungungen ermöglichen es Spektrale Einhüllende so zu berechnen, dass die Frequenzauflösung des menschlichen Gehörs nachgebildet werden kann. Berechnungen der Spektralen Einhüllenden auf Basis der *Linearen Prädiktion* (LP) mit Mel-Frequenzanpassung haben im Bereich der Spracherkennung und -codierung gegenüber der herkömmlichen LP eindrucksvolle Verbesserungen gebracht. Allerdings birgt die LP das Problem, bei stimmhafter Sprache mit mittlerer oder hoher Tonlage, die für die Spracherkennung sehr wichtigen spektralen Spitzen zu überzeichnen. Murthi und Rao bemerkten, dass durch Ersetzung der LP durch die *Minimum Variance Distortionless Response* (MVDR) hoher Ordnung das Problem der Überzeichnung behoben werden kann und haben somit die Möglichkeit geschaffen, die Einhüllenden den Spektren besser anzupassen. Um die Stärken beider Ansätze, der Mel-Frequenzanpassung des menschlichen Gehörs und der genaueren Modellierung zu vereinen, schlagen wir vor die Mel-Frequenzanpassung in das MVDR-All-Pol-Modell zu integrieren und adaptieren ein Verfahren zur schnellen Berechnung.

Im Vergleich mit der Fouriertransformation, der LP, der MVDR und der Mel-LP in der akustischen Vorverarbeitung des Automatischen Spracherkenners *Janus Recognition Toolkit* (JRTk), entwickelt von der Universität Karlsruhe und der Carnegie Mellon University, konnte unser Ansatz, die Mel-MVDR, ohne Sprechernormierung bereits überzeugen [11]. In dieser Arbeit zeigen wir, dass die Mel-MVDR auch mit Vokaltraktlängennormierung Verbesserungen in der Spracherkennung liefert. Zusätzlich untersuchen wir inwieweit die Bilineare Transformation zur Sprechernormierung geeignet ist. Die Ergebnisse sind tabellarisch zusammengestellt und werden diskutiert.

## 1 Einleitung

Im Gegensatz zum menschlichen Gehör, das in niedrigeren Frequenzen eine höhere Auflösung als in hohen Frequenzen bereitstellt, löst ein All-Pol Modell oder eine Fouriertransformation alle Frequenzen gleich gut auf. Ein gängiger Ansatz in der automatischen Spracherkennung die Auflösung des menschlichen Gehöres nachzuahmen besteht darin die über ein All-Pol Modell gewonnene Einhüllende oder das durch die Fouriertransformation gewonnene Spektrum mit einer Mel-Filterbank nachzubehandeln, hierzu siehe auch Abbildung 2 links. Hierbei ist zu beachten, dass keine Erhöhung der Auflösung der Einhüllenden bei niedrigen Frequenzen

erzielt werden kann. Um dies zu erreichen müssen die bisher gleich verteilten All-Pole zur Berechnung der Einhüllenden neu verteilt werden, so dass mehr Pole zur Beschreibung der niedrigen Frequenzen, als der hohen, zur Verfügung stehen. Strube [8] schlug vor dies durch eine Dehnung[1] der Frequenzachse zu erreichen, ermöglicht durch die *Bilineare Transformation*, und wendete dies auf die *Lineare Prädiktion* (LP) an. Wird der Dehnfaktor so gewählt, dass er die Mel-Frequenz annähert, kommt es bei männlichen Sprechern zu einer signifikanten Reduktion der Wortfehlerrate im Vergleich zur herkömmlichen LP und einer leichten Reduktion zu den weit verbreiteten *Mel-Frequenz Kepstral Koeffizienten* (MFKK) [4]. Für weibliche Sprecher ist gegenüber der Mel-LP eine Verbesserung möglich, aber nicht gegenüber den MFKK. Eine Erklärung hierfür ist, dass die Spektrale Einhüllende, gewonnen durch LP, weit auseinadergezogene harmonische Spitzen [6] überhöht wie sie insbesondere bei weiblichen Sprechern vorkommen, da sie eine höhere Fundamentalfrequenz besitzen. Murthi und Rao haben bemerkt, dass dieses Problem durch Verwendung einer Spektralen Einhüllenden, berechnet durch das *Minimum Variance Distortionless Response* (MVDR) All-Pol Modell, überwunden werden kann [6].

Um die Vorzüge der gesteigerten Frequenzauflösung ähnlich des menschlichen, auditiven Systems mit der besseren Approximation der Spektralen Einhüllenden zu kombinieren schlagen wir vor die Mel-Frequenzanpassung in die MVDR-Berechnung zu integrieren, die wir in Anlehnung an Mel-LP, Mel-MVDR All-Pol Modell nennen möchten. Des weiteren ermöglicht eine Adaption des von Musicus [7] vorgeschlagen Algorithmus eine schnelle Berechnung auf Basis der Mel-LP Koeffizienten.

## 2 Theoretischer Hintergrund

Beim Schätzen der MVDR[2] Einhüllenden ist zu berücksichtigen, dass das Signal an der Frequenz $\omega_l$ verzerrungsfrei übertragen wird [2]. D.h. für die Impulsantwort muss gelten:

$$H(e^{j\omega_l}) = \sum_{k=0}^{M} h^*(k) e^{-jk\omega_l} = 1$$

Mit dem Frequenzvektor

$$\mathbf{s}(\omega) = [1, e^{-j\omega}, \ldots, e^{-jM\omega}]^T \tag{1}$$

und $\mathbf{h} = [h(0), h(1), \ldots, h(M)]^T$ kann hierfür vereinfacht geschrieben werden:

$$\mathbf{s}^H(\omega_l) \cdot \mathbf{h}^* = 1$$

Um die Frequenz nichtlinear abbilden zu können ersetzen wir die Verzögerungselemente $e^{-jk\omega}$ des Frequenzvektors $\mathbf{s}(\omega)$ durch All-Pass Selektion, z.B. mit dem *All-Pass Filter erster Ordnung*, besser bekannt als *Bilineare Transformation*:

$$e^{-j\tilde{\omega}} = D(e^{-j\omega}) = \frac{e^{-j\omega} - \alpha}{1 - \alpha \cdot e^{-j\omega}} \tag{2}$$

Hierbei ist $\alpha$ der *Dehnfaktor* und $D(e^{-j\omega})$ das *gedehnte Verzögerungsglied*. Die Phasenfunktion, bzw. die Frequenzabbildungungsfunktion, von $D(e^{-j\omega})$ läßt sich berechnen zu [4]:

$$\arg\left(D(e^{-j\omega})\right) = \tilde{\omega} = \omega + 2\arctan\frac{\lambda \sin\omega}{1 - \lambda\cos\omega}$$

[1] Dehnung soll hier sowohl als ein Auseinanderziehen als auch Zusammendrücken verstanden werden.

[2] Die MVDR wurde zuerst von Capon [1] vorgestellt und ist auch bekannt als Maximum-Likelihood Methode.

D.h. die lineare Frequenzachse wird ungleichmäßig gedehnt. Durch geschickte Wahl des Dehnfaktors läßt sich die Mel-Frequenz, als auch die Bark-Frequenz, nachahmen, gezeigt in Abbildung 1. Hierbei ist zu beachten, dass sich der Dehnfaktor in Abhängigkeit von der Abtastfrequenz ändert.

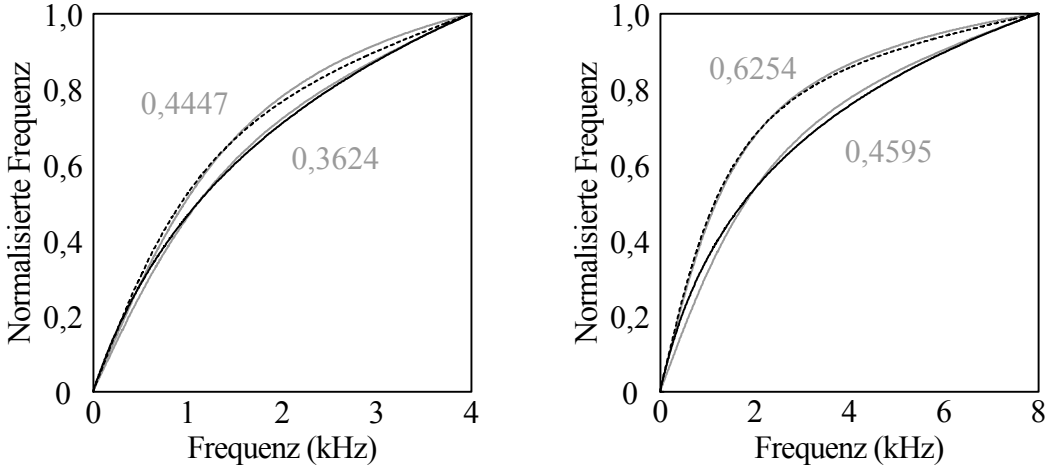


**Abbildung 1** - Gezeigt wird die Approximation der Bilinearen Transformation (graue Linien einschließlich der Dehnfaktoren als graue Zahlen) an die Mel-Frequenz (schwarze Linien) und die Bark-Frequenz (gepunktete schwarze Linien) für Abtastraten von 8 kHz (links) und 16 kHz (rechts).

Durch Einsetzen des All-Pass Filters erster Ordnung (2) in den Frequenzvektor (1) ergibt sich der *gedehnte Frequenzvektor*:

$$\tilde{\mathbf{s}}(\omega) = \left[1, \frac{e^{-j\omega} - \alpha}{1 - \alpha \cdot e^{-j\omega}}, \ldots, \frac{e^{-jM\omega} - \alpha}{1 - \alpha \cdot e^{-jM\omega}}\right]^T \tag{3}$$

Der Filter $\mathbf{h}_\mathrm{l}$ kann somit berechnet werden durch das *gedehnte Minimierungsproblem* das die Ausgangsenergie des gesamten, gedehnten Frequenzbereiches minimiert:

$$\min_{\tilde{\mathbf{h}}_\mathrm{l}} \tilde{\mathbf{h}}_\mathrm{l}^H \phi \tilde{\mathbf{h}}_\mathrm{l} \;\text{ unter der Bedingung, dass }\; \tilde{\mathbf{s}}^H(\omega_\mathrm{l})\tilde{\mathbf{h}}_\mathrm{l} = 1$$

hierbei ist $\phi$ die $(M+1)\cdot(M+1)$ Toeplitz Autokorrelationsmatrix des Filtereingangs. Die Lösung des gebeugten Minimierungsproblems ist ähnlich ihres ungedehnten Gegenstücks, dessen Herleitung in [2] ausgeführt ist. Man beachte, dass hier der gedehnte Frequenzvektor $\tilde{\mathbf{s}}$ anstelle des Frequenzvektor $\mathbf{s}$ eingesetzt ist.:

$$\tilde{\mathbf{h}}_\mathrm{l} = \frac{\phi^{-1}\tilde{\mathbf{s}}(\omega_l)}{\tilde{\mathbf{s}}^H(\omega_l)\phi^{-1}\tilde{\mathbf{s}}(\omega_l)}$$

D.h. die Impulsantwort des verzerrungsfreien Filters für die Frequenz $\omega_l$ ist gegeben durch $\tilde{\mathbf{h}_l}$. Die gedehnte MVDR Spektralschätzung berechnet sich hiermit zu

$$\tilde{S}_{\mathrm{MVDR}}(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\tilde{\mathbf{H}}_\mathrm{l}(e^{j\omega})\right|^2 S(e^{-j\omega})d\omega$$

wobei $S(e^{-j\omega})$ die Signalenergie bei der Frequenz $\omega_l$ ist. Die verzerrungsfreie Übertragung bei der Frequenz $\omega_l$ gemeinsam mit der Minimierung der Ausgangsenergie garantieren, dass die

verbleibenden Frequenzkomponenten in idealer Weise unterdrückt werden. Die Berechnung einzelner Filter für beliebige Frequenzen $\omega_l$ ist nur konzeptionell. Tatsächlich kann gezeigt werden, dass das MVDR Spektrum über alle Frequenzen hinweg berechnet werden kann [2], was auch auf das gedehnte MVDR Spektrum zutrifft:

$$\tilde{S}_{\mathrm{MVDR}}(\omega) = \frac{1}{\tilde{\mathbf{s}}^H(\omega)\phi^{-1}\tilde{\mathbf{s}}(\omega)}$$

Mit der Annahme, dass die Toeplitz Autokorrelationsmatrix positiv definit und folglich invertierbar ist, ist es möglich einen schnellen Algorithmus aufzustellen mit dem die MVDR aus den LP Koeffizienten entwickelt werden kann [7]. Da zwischen den gedehnten MVDR Koeffizienten und den gedehnten LP Koeffizienten die gleiche Beziehung besteht wie zwischen den ungedehnten MVDR Koeffizienten zu den ungedehnten LP Koeffizienten kann der Algorithmus in leicht abgeänderter Form verwendet werden:

1. **Berechnung der gedehnten LP Koeffizienten**
   Zur Berechnung der gedehnten LP Koeffizienten gibt es verschiedene Möglichkeiten. Bei unseren Versuchen wurde ein Algorithmus wie von Matsumoto u.a. vorgeschlagen verwendet. [4].

2. **Korrelation der gedehnten Vorhersagekoeffizienten**

$$\tilde{\mu}_k = \begin{cases} \sum_{i=0}^{N-k}(N+1-k-2i)\tilde{a}_i^{(N)}\tilde{a}_{i+k}^{*(N)} & : k = 0, \cdots, N \\ \tilde{\mu}_{-k}^* & : k = -N, \cdots, -1 \end{cases}$$

3. **Berechnung des gedehnten MVDR Spektrums**

$$\tilde{S}_{\mathrm{MVDR}}(\omega) = \frac{\epsilon}{\sum_{k=-M}^{M}\tilde{\mu}_k e^{-j\omega k}} \tag{4}$$

Hier sei angemerkt, dass das Spektrum wie in (4) berechnet bereits das Mel-approximierte Spektrum abbilden kann und somit die Mel-Filterbank durch eine Lineare-Filterbank von halb überlappenden Dreiecken oder besser noch durch eine Adaptierte-Filterbank, um die Differenz zwischen dem Mel-Spektrum und der Bilinearen Transformation auszugleichen, ersetzt werden kann, vergleiche hierzu die verschiedenen Filterbänke in Abbildung 2

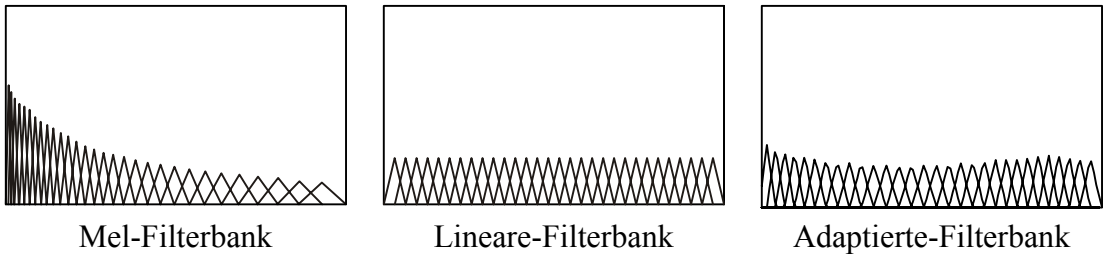


**Abbildung 2** - Verschiedene Filterbänke im Vergleich.

Die gedehnte Einhüllende unterscheidet sich primär in der Verteilung der Parameter die zur Beschreibung der Einhüllenden verwendet wird. Während die ungedehnte MVDR die Parameter gleich verteilt, werden bei der gedehnten MVDR mehr Parameter zur Beschreibung der niedrigen Frequenzen verwendet und entsprechend weniger in den hohen Frequenzen.

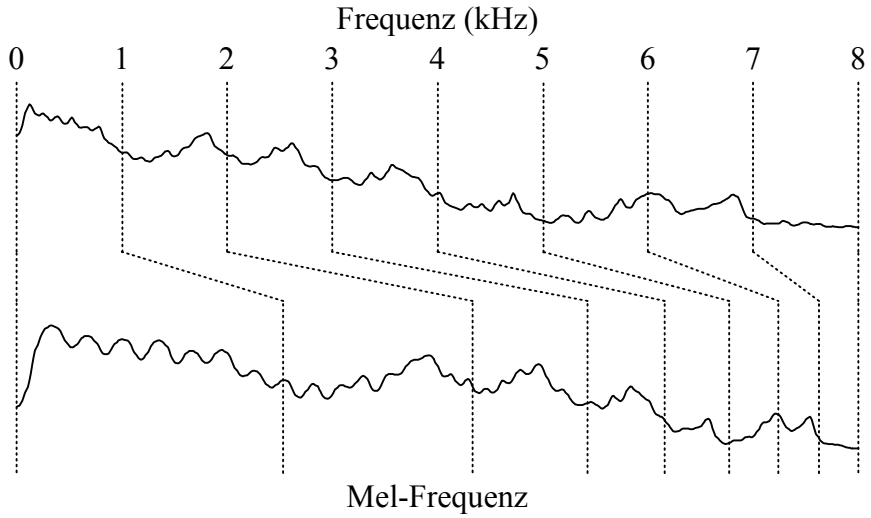


**Abbildung 3** - Vergleich der MVDR Einhüllenden (oben) und der Mel-MVDR Einhüllenden (unten), beide mit Modellordnung 120.

Abbildung 3 illustriert den Unterschied zwischen den MVDR und Mel-MVDR Einhüllenden. Hierbei ist der Dehnfaktor auf 0,4595 gesetzt um die Mel-Frequenz für ein mit 16 kHz abgetastetes Signal zu simulieren. Hier ist zu sehen wie die gedehnte MVDR eine höhere Auflösung für Frequenzen bis 2 kHz bereitstellt mit immer weiter fallender Auflösung hin zu höheren Frequenzen. Die gedehnte MVDR stellt somit eine interessante Eigenschaft zur Verfügung, die durch die herkömmliche Verwendung der MVDR nicht bereit gestellt werden kann:
Die Residien zeigen spektrale Abflachungen ähnlich der Feuerrate des auditiven Nerves. Dies ist ähnlich dem Ergebnis der gedehnten LP [3], aber ohne den negativen Effekt der Überzeichnung der harmonischen Spitzen, wie es für die LP bei Sprache mittlerer und hoher Frequenz typisch ist .

## 3 Anwendung in der automatischen Spracherkennung

Die unten aufgeführten Spracherkennerexperimente wurden mit dem *Janus Recognition Toolkit* (JRTk) durchgeführt, das gemeinsam von den Schwesterlaboratorien, den *Interactive Systems Laboratories*, an der Universität Karlsruhe (TH), Deutschland und an der Carnegie Mellon University in Pittsburgh, Pennsylvania, USA entwickelt und gepflegt wird.

Alle Experimente wurden mit dem *Switchboard Corpus* durchgeführt. Hierbei handelt es sich um kontinuierliche englische Mensch zu Mensch Kommunikation, aufgenommen über analoge Telefonleitungen und analoge Funkverbindungen im Nordamerikanischen Kontinent. Es wurden 548 Sprecher beiden Geschlechts zum Training und 16 Sprecher beiden Geschlechts für den Test verwendet, wovon zwei Gespräche über Handies geführt wurden.

Verwendet wurden 4.166 Kodebücher beschrieben durch je 32 Gaussiane, was sich zu einer Gesamtzahl von 133.312 Gaussiane berechnet. Die Merkmale wurden alle 10 ms neu berechnet wobei das mit 8 kHz abgetastete, kontinuierliche Sprachsignal durch ein 20 ms Hamming Fenster zerlegt wurde. Um die von uns vorgeschlagene Methode zu vergleichen berechnen wir 13 Kepstralkoeffizienten, zusammen mit ihren ersten und zweiten Ableitungen, durch eine diskrete Kosinustransformation unter Verwendung verschiedener Spektralrepräsentationen:

- Der *Schnellen Fouriertransformation* (FFT) und der MVDR, beide gefolgt von einer Mel-Filterbank, bestehend aus 30 halb überlappenden Mel-verteilten triangulären Filtern, Abbildung 2 links.
- Der Mel-MVDR gefolgt von 30 identischen, halb überlappenden gleich verteilten triangulären Filtern, Abbildung 2 Mitte.
- Der Mel-MVDR gefolgt von 30 überlappenden, so verteilten Filtern, dass die Differenz der Bilinearen Transformation zur Mel-Frequenz ausgeglichen wird, Abbildung 2 rechts.

Um eine gute Vergleichsmöglichkeit der verschiedenen Methoden zu gewährleisten werden alle spektralen Einhüllenden rekonstruiert und die höchste Energie auf den höchsten Energiebereich der Fouriertransformation skaliert [10]. Um die Kanalvariationen zu kompensieren wird die Kepstrale Mittelwertsubtraktion verwendet. Die Merkmale werden zu ihrer endgültigen Anzahl, 32, durch eine Lineare Diskriminanzanalyse reduziert.

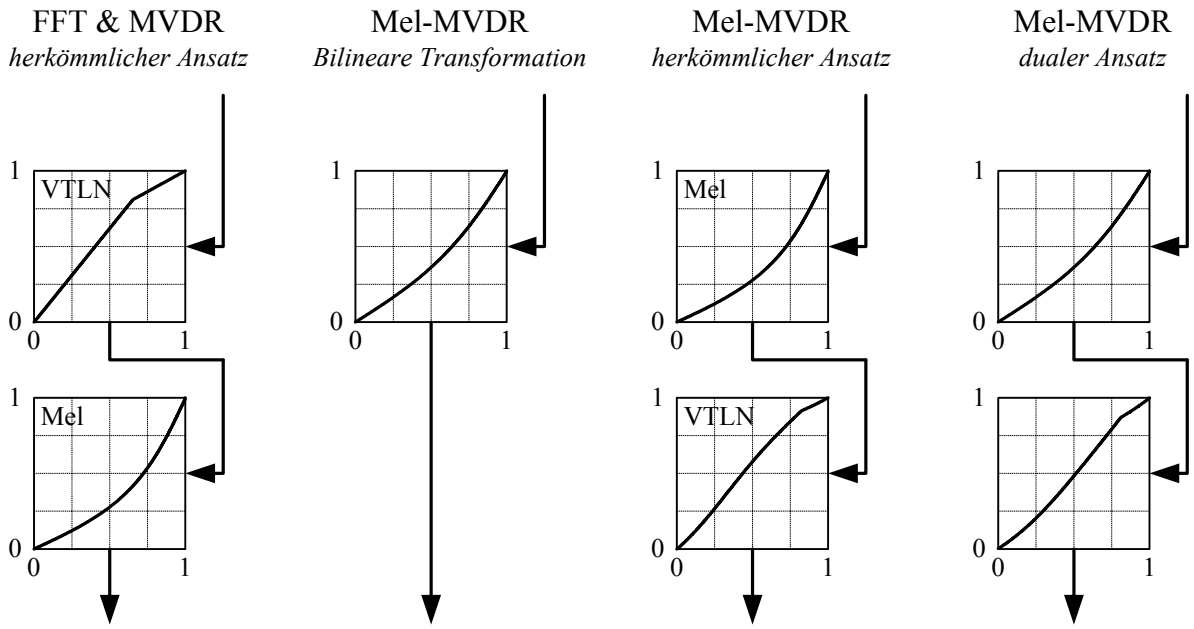


**Abbildung 4** - Vergleich verschiedener Frequenzabbildungen (alle auf eins normiert) zur Implementation von VTLN und Mel-Frequenz. Es sei angemerkt, dass sowohl bei der Bilinearen Transformation als auch dem dualen Ansatz in allen Blöcken sowohl VTLN als auch Mel-Anpassung durchgeführt wird.

Es sei darauf aufmerksam gemacht, dass die *Vokaltraktlängennormierung* (VTLN) auf unterschiedliche Weise erfolgt. Während bei der Fouriertransformation und der MVDR Einhüllenden die Anpassung im linearen Frequenzbereich erfolgt, oberer Block in Abbildung 4 links, untersuchen wir bei der Mel-MVDR drei verschiedene Ansätze:

- **Bilineare Transformation**
  Hier wird ausschließlich die Bilineare Transformation zur VTLN verwendet, nur ein Block siehe Abbildung 4 Mitte links.
- **Herkömmlicher Ansatz**
  Der herkömmliche Ansatz basiert auf der linearen Verzerrung [9], allerdings im

gebeugten Frequenzbereich, siehe unterer Block in Abbildung 4 Mitte rechts. Um dies zu ermöglichen wird die lineare Verzerrung auf den gebeugten Bereich umgerechnet und ist somit nicht mehr linear, weshalb bewußt auf den Begriff der Linearität verzichtet wird.

- **Dualer Ansatz**
  Im dualen Ansatz wird primär die Bilineare Transformation zur VTLN verwendet, aber zusätzlich werden die Filterbänke so berechnet, dass sie den Unterschied zwischen der Bilinearen Transformation und dem herkömmlichen Ansatz ausgleichen, beide Blöcke in Abbildung 4 rechts.

## 4 Diskussion

Mit den hier vorgestellten Experimenten, Tabelle 1, kann bestätigt werden, dass die vorgeschlagene Mel-Frequenzanpassung der MVDR sowohl gegenüber der Fouriertransformation als auch gegenüber der MVDR eine Verbesserung der Erkennergenauigkeit liefert. Dies gilt sowohl für den Fall ohne VTLN als auch für den Fall mit VTLN. Zur VTLN bewährt sich der herkömmliche Ansatz gegenüber dem Einsatz der Bilinearen Transformation und des dualen Ansatzes.

| ***ohne Vokaltraktlängennormierung*** | | ***mit Vokaltraktlängennormierung*** | |
|---|---|---|---|
| FFT<br>*Mel Filterbank* | 39,1 | FFT<br>*herkömmlicher Ansatz* | 37,2 |
| MVDR (80)<br>*Mel Filterbank* | 38,7 | MVDR (80)<br>*herkömmlicher Ansatz* | 37,0 |
| Mel-MVDR (50)<br>*lineare Filterbank* | 38,4 | Mel-MVDR (50)<br>*Bilineare Transformation* | 37,0 |
| Mel-MVDR (50)<br>*adaptierte Filterbank* | 38,5 | Mel-MVDR (50)<br>*herkömmlicher Ansatz* | 35,9 |
| | | Mel-MVDR (50)<br>*dualer Ansatz* | 36,3 |

**Tabelle 1** - Dargestellt sind die Wortfehlerrraten ohne (linke Spalte) und mit (rechte Spalte) VTLN für verschiedene, akustische Vorverarbeitungen.

Allgemein kann ein möglicher Gewinn spektraler Einhüllenden gegenüber dem Fourierspektrum in der Art und Weise erklärt werden in der spektrale Täler (d.h. Bereiche mit geringer Energie) und Spitzen (d.h. Bereiche mit hoher Energie) modelliert werden:
Während die Fouriertransformation spektrale Täler und Spitzen mit gleicher spektraler Auflösung modelliert, beschränken sich spektrale Einhüllende auf eine akurate Beschreibung der spektralen Spitzen. Spektrale Täler werden überdeckt, d.h. sie werden nur ungenau nachgebildet, und somit kommt es in diesen Bereichen zur Verringerung der Varianz gegenüber Störgeräuschen. Dies ist insbesondere wertvoll, da sich Geräusche, im logarithmischen Spektrum, primär in Frequenzen mit geringer Energie störend auf die Akkuratheit eines Spracherkenners auswirken, wohingegen Frequenzen mit hoher Energie kaum gestört werden.

Es sei angemerkt, dass die Verwendung der Burg-Frequenz an Stelle der Mel-Frequenz zu einer Verschlechterung der Erkennergenauigkeit führt weshalb sich unsere Untersuchung nur auf die Mel-Frequenz beschränkte.

Im Gegensatz zur LP Einhüllenden bewegen sich bei der MVDR Einhüllenden die Formanten nicht in Abhängigkeit von der Modellordnung[6] womit sich zukünftige Aktivitäten auf eine variable Modellordnung in Abhängigkeit von Sprecher, Signal zu Rauschverhältnis oder Maximum Likelihood konzentrieren könnten um eine bessere Anpassung an die gelernten Modelle zu erreichen.

## Literatur


[1] Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE*, vol. 57:pp. 1408–1418, August 1969.

[2] Haykin, S. *Adaptive filter theory—3th ed.* Prentice Hall, 1991.

[3] Karjalainen, M. Auditory interpretation and application of warped linear prediction. *Proceedings of Consistent & Reliable Acoustic Cues for Sound Analysis*, 2001.

[4] Matsumoto, H. und Moroto, M. Evaluation of Mel-LPC cepstrum in a large vocabulary continuous speech recognition. *IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 1:pp. 117–120, 2001.

[5] McDonough, J.W. *Speaker compensation with all-pass transforms.* Ph.D. thesis, Johns Hopkins University, Baltimore, USA, 2000.

[6] Murthi, M.N. und Rao, B.D. All-pole modeling of speech based on the minimum variance distortionless response spectrum. *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2000

[7] Musicus, B.R. Fast MLM power spectrum estimation from uniformly spaced correlations. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33:pp. 1333–1335, 1985.

[8] Strube, H.W. Linear prediction on a warped frequency scale. *Journal Acoustic Society of America*, vol. 68(no. 8):pp. 1071–1076, 1980.

[9] Westphal, M. *Robuste kontinuierliche Spracherkennung für mobile Informationssysteme.* Doktorarbeit, Universität Karlsruhe (TH), Karlsruhe, Deutschland, Juni 2000.

[10] Wölfel, M.C. *Minimum variance distortionless response spectral estimation and subtraction for robust speech recognition.* Diplomarbeit, Universität Karlsruhe (TH), Karlsruhe, Deutschland, Januar 2003.

[11] Wölfel, M.C.; McDonough, J.W. und Waibel, A. Minimum variance distortionless response on a warped frequency scale. *Eurospeech 2003.*