Optimierung des Multiband-Excitation-(MBE)-Verfahrens für den Einsatz in der Sprachsynthese

Ulrich Kordon, Wolfram Manthey
Technische Universität Dresden, Institut für Akustik und Sprachkommunikation

1. Einführung

Moderne Text-to-Speech-(TTS)-Sprachsynthesesysteme arbeiten mit Spracheinheiteninventaren auf Zeitfunktionsbasis, wobei zunehmend größere Einheiten (z.B. Silben oder Wörter) verwendet werden. Das damit verbundene Anwachsen des Speicherbedarfs für das Spracheinheiteninventar sowie Einschränkungen herkömmlicher Verfahren bei prosodischen Manipulationen direkt in der Zeitfunktion erfordern den Einsatz entsprechender Präsentationsverfahren für das Spracheinheiteninventar. Neben den aus der Audiosignalcodierung bekannten perzeptiven Verfahren bieten sich dafür Verfahren aus der nachrichtentechnischen Sprachcodierung an. Als besonders leistungsfähig hat sich dabei das Multiband-Excitation-(MBE)-Verfahren [1] herausgestellt, bei dem auf der Basis eines konventionellen Anregungs-Bewertungsmodells eine spektralabhängige Mischung aus stimmhaftem und stimmlosem Anregungssignal verwendet wird.
Nach einem Überblick zu MBE-Grundprinzip und darauf basierenden konkreten Verfahren wird der Einsatz des in der Sprachcodierung verwendeten MBE-Grundverfahrens in einem TTS-Sprachsynthesesystem näher untersucht. Der Schwerpunkt liegt dabei auf einem Vergleich verschiedener Möglichkeiten zur Ermittlung der MBE-Parameter der Spracheinheiten und Aussagen zur Leistungsfähigkeit bei prosodischen Signalmanipulationen.

2. Prinzip und Verfahren

Das MBE-Verfahren basiert prinzipiell auf einem spektralen Anregungs-Bewertungsmodell. Das Kurzzeitspektrum $S(\omega)$ eines Signalabschnitts soll dabei durch ein Modellspektrum $S_M(\omega)$ als Produkt aus einem Anregungsanteil $E(\omega)$ und einem Bewertungsanteil $H(\omega)$ angenähert werden:

$$S(\omega) \approx S_M(\omega) = E(\omega) \cdot H(\omega) \tag{1}$$

Die Schätzung des Bewertungsanteils $H(\omega)$ bei gegebenem $S(\omega)$ kann z.B. durch eine lineare Interpolation der Spektrallinien oder durch Linear-Prediction- bzw. Cepstralverfahren erfolgen.
Der Anregungsanteil $E(\omega)$ wird beim MBE-Ansatz durch eine gewichtete additive Überlagerung einer periodischen Komponente $E_{per}(\omega)$ und einer Rauschkomponente $E_{rausch}(\omega)$ gebildet:

$$E(\omega) = a(\omega) \cdot E_{per}(\omega) + b(\omega) \cdot E_{rausch}(\omega) \tag{2}$$

Der periodische Anteil kann durch eine Pulsfolge mit der Grundfrequenz des Originalsignals und flacher spektraler Hülle realisiert werden. Als Rauschkomponente dient weißes Rauschen. Die frequenzabhängigen Wichtungsfaktoren $a(\omega)$ und $b(\omega)$ werden nun so bestimmt, dass sich auf der Basis eines Fehlerkriteriums minimale Abweichungen zwischen $S(\omega)$ und $S_M(\omega)$ ergeben. Der jeweilige Signalabschnitt wird dann durch die MBE-Modellgrößen $H(\omega)$,$a(\omega)$ sowie $b(\omega)$ repräsentiert
Basierend auf diesem Grundprinzip wurden verschiedene Verfahren entwickelt. Sie zielen vor allem auf eine effektivere Präsentation der spektralen Hülle, eine Erhöhung Sprachqualität sowie eine optimale Anpassung an die Haupteinsatzgebiete Sprachcodierung bzw. Sprachsynthese. Die Präsentation der spektralen Hülle kann nichtparametrisch, d.h. in Form von Stützstellen, oder parametrisch erfolgen. Für die parametrischen Verfahren kommen vor allem Linear-Prediction-Modelle zum Einsatz (z.B. Reflexionskoeffizienten, Line-Spektral-Pairs). Die Wichtungsfaktoren werden für bestimmte Frequenzbereiche angegeben, wobei hier lineare (z.B. grundfrequenzsynchron) oder gehörgerechte (z.B. critical bands) Anordnungen verwendet werden. Durch Berücksichtigung der Phaseninformationen ist eine Steigerung der Sprachqualität möglich. Für den Einsatz in Sprachsynthesesystemen wurden kombinierte Verfahren entwickelt, die bezüglich Aufwand und erreichbarer Sprachqualität für diesen Anwendungsbereich ein Optimum darstellen [2]. Die MBE-Verfahren werden in der Sprachcodierung vor allem für geringe Bitraten im Bereich um 2,4 kBit/s eingesetzt. Durch Kombination mit Vektorquantisierungsverfahren sind Datenraten unter 1kBit/s möglich.

3. Einsatz des MBE-Verfahrens in der Sprachsynthese

3.1. Anforderungen

Beim Einsatz in Sprachsynthesesystemen ergeben sich einige spezifische Anforderungen an das Präsentationsverfahren, die aus den Besonderheiten der dort verwendeten Spracheinheiteninventare resultieren:
- Die Spracheinheiten stellen unter Umständen sehr kurze Abschnitte dar, so dass keine Adaptionszeiten zur Verfügung stehen.
- Ihre Verkettbarkeit darf durch das Präsentationsverfahren nicht beeinträchtigt werden.
- Die Variation von Grundfrequenz und Dauer muss störungsfrei über Bereiche von ca. 0,5 bis 3 möglich sein.
- Es darf möglichst kein hörbarer Qualitätsverlust auftreten.
- Die gewünschten Datenreduktionsraten liegen im Bereich 10 ... 50.

3.2. Realisiertes Verfahren und Optimierung

Prinzipiell kann für die Spracheinheitenpräsentation auf Standardverfahren zurückgegriffen werden, wie sie z.B. in der Sprachcodierung Verwendung finden. Bei verschiedenen Anwendungen stehen entsprechende Module ohnehin zur Verfügung (z.B. Telekommunikationsanwendungen). Deshalb soll hier untersucht werden, inwieweit und unter welchen Bedingungen das MBE-Grundverfahren den in Abschnitt 3.1. aufgeführten speziellen Anforderungen genügt.
Als Sprachsynthesesystem wurde das Dresdener Sprachsynthesesystem DRESS verwendet [3]. In Bild 1 ist der realisierte Modul zur Spracheinheitenpräsentation auf MBE-Basis [4] dargestellt.

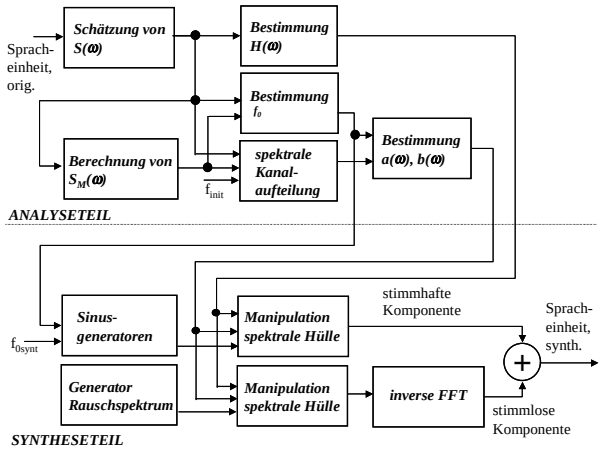

Bild 1 Realisiertes MBE-System zur Präsentation von Spracheinheiten

Im Analyseteil werden die das Inventar von DRESS bildenden ca. 1200 Diphon-Zeitfunktionen zunächst zur Bestimmung der $S(\omega)$ und $S_M(\omega)$ einer spektralen Kurzzeitanalyse unterzogen. Daran schließt sich die Extraktion der jeweiligen MBE-Parameter spektrale Hülle $H(\omega)$, Wichtungsfaktoren $a(\omega)$, $b(\omega)$ sowie Grundfrequenz f_0 bei

stimmhaften Abschnitten an. *a(ω)* und *b(ω)* werden nun so gewählt, dass sich für den jeweiligen quadratischen Fehler ein Minimum ergibt

$$\varepsilon = \frac{1}{2\pi}\int_{-\pi}^{+\pi}(|S(\omega)|-|S_M(\omega)|)^2 d\omega \qquad (3)$$

Die Synthese erfolgt durch Addition einer stimmhaften und stimmlosen Signalkomponente, wobei die stimmhaft-Komponente durch additive Überlagerung von mit a*(ω)H(ω)* gewichteten Sinussignalen der Frequenzen $n \cdot 2\pi \cdot f_{0synt}$ gebildet wird ($1 \le n \le f_{abtast}/2 \cdot f_{0synt}$, f_{abtast} ... Abtastfrequenz), während die stimmlos-Komponente durch Fourierrücktransformation eines mit *b(ω)H(ω)* gewichteten weißen Rauschspektrums entsteht.
Um das realisierte MBE-Verfahren bezüglich der in Abschnitt 3.1. formulierten Anforderungen zu optimieren, wurden verschiedene Varianten von *H(ω)* und *a(ω)*/*b(ω)* verglichen und einer subjektiven Bewertung unterzogen.

4. Ergebnisse

In Bild 2 sind exemplarisch 4 verschiedene Varianten zur Ermittlung von *H(ω)* gegenübergestellt. Aus der Darstellung geht hervor, dass die lineare Interpolation der cepstralen Hülle (80 Koeffizienten) das beste Abbildungsverhalten zeigt.

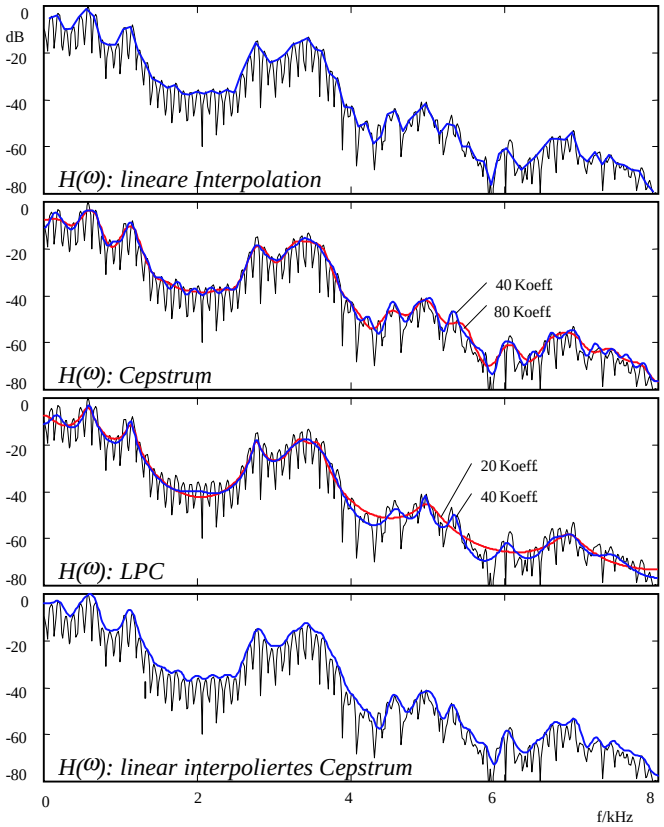

Bild 2 Vergleich von Varianten zur Ermittlung von H(ω)

Zur Beurteilung der erreichbaren Gesamtqualität wurden zunächst originale Sprachsignale MBE-codiert und wieder resynthetisiert. Für *a(ω)* und *b(ω)* kamen binäre Funktionen mit $a(\omega)+b(\omega)=1$ und eine Kanalaufteilung nach Frequenzgruppen (22 Kanäle) zum Einsatz . Diese Tests dienten außerdem zur Ermittlung des prosodischen Variationsbereichs, wobei die Forderung im Abschnitt 3.1. erfüllt wird. Bild 3 stellt die Sonagramme der von DRESS synthetisierten Äußerung „Guten Tag“ mit Time-Domain-(TD)-PSOLA und der als optimal gefundenen MBE-Variante (linear interpolierte cepstrale Hülle, Frequenzgruppen-Kanalaufteilung) gegenüber.

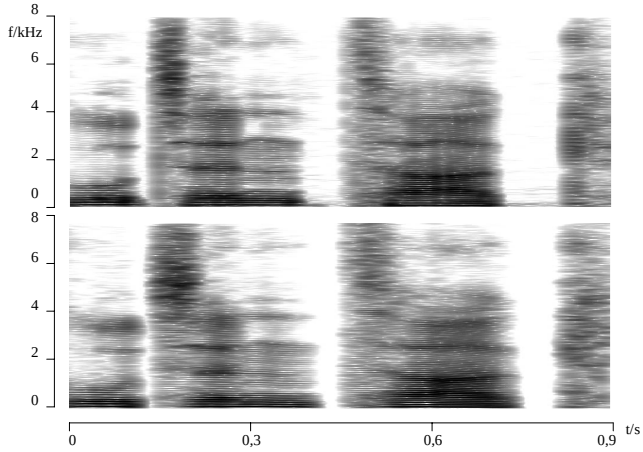

Bild 3 Synthetisierte Äußerung „Guten Tag“ (oben Time-Domain-PSOLA, unten MBE)

Die subjektive Bewertung lieferte für die optimale Variante (lineare Interpolation der cepstralen Hülle) einen MOS (1 schlecht, 5 ausgezeichnet) von 2,36 (äquivalent Time-Domain-PSOLA), wobei der Datenreduktionsfaktor bei 1,9 lag. Er konnte bis ca. 3,5 erhöht werden (Verringerung der Parameterzahl von *H(ω)* auf 40), ohne dass ein nennenswerter Qualitätsverlust auftrat (MOS 2,23).

5. Zusammenfassung

Das MBE-Grundverfahren stellt eine Alternative zum Time-Domain-PSOLA-Verfahren dar. Als optimale Verfahrensvariante bezüglich Qualität und Aufwand kann die lineare Interpolation der cepstralen Hülle bei spektraler Kanalaufteilung nach Frequenzgruppen und binären Wichtungsfunktionen angesehen werden. Vorteile gegenüber TD-PSOLA besitzt das MBE-Verfahren vor allem bei der Manipulation der prosodischen Parameter und durch die Reduktion des Speicherbedarfs für das Inventar. Allerdings steht dem der erhöhte algorithmische Aufwand und im Vergleich zu z.B. perzeptiven Verfahren [5] relativ geringe Datenreduktionsfaktor gegenüber.

6. Literatur

[1] Griffin, D. W., Lim, J. S.: Multiband Excitation Vocoder. IEEE Transactions on Acoustics, Speech and Signal Processing, 36(1988)8, S. 1223 – 1235.

[2] Dutoit, T., Leich, H.: Improving the TD-PSOLA Text-to-Speech Synthesizer with a Specially Designed MBE Re-Synthesis of the Segments Database. In: Vandewalle, J., Boite, R., Moonen, M., Oosterlinck, A. (eds.): Signal Processing VI: Theories and Applications, Elsevier Science Publisher B. V., 1992.

[3] Hoffmann, R.: A multilingual text-to-speech system. The Phonetician 80 (1999 / II), p. 5 – 10.

[4] Manthey, W.: Multiband-Anregungsverfahren für die Signalerzeugung in Sprachsynthesesystemen. Diplomarbeit, TU Dresden, Institut für Akustik und Sprachkommunikation, 2000.

[5] Kordon, U.: Möglichkeiten zur Datenreduktion von Spracheinheiteninventaren für die Sprachsynthese. KONVENS 2000, Sprachkommunikation, ITG-Fachbericht 161, VDE Verlag Berlin – Offenbach, S. 255 – 258.