Modellbasierte Analyse und Verkettung von Diphonen für die Spracherzeugung

Karl Schnell, Arild Lacroix

Institut für Angewandte Physik, Goethe-Universität Frankfurt
Max-von-Laue-Straße 1, D-60438 Frankfurt am Main,
Email: {Schnell, Lacroix}@iap.uni-frankfurt.de

Einleitung

Die Synthese von beliebigen Sprachäußerungen wird gewöhnlich durch eine Verkettung von Spracheinheiten vorgenommen. Als Spracheinheiten werden in diesem Beitrag Diphone verwendet, die mit Hilfe des verlustbehafteten Rohrmodells analysiert und synthetisiert werden. Ein Problem des Verkettungsansatzes besteht darin, daß an den Verkettungsstellen der Diphone Unstetigkeiten auftreten. Es wird gezeigt, wie die auftretenden Unstetigkeiten durch eine angepasste Analyse der verwendeten Diphondatenbank verringert werden.

Sprechtraktmodell

Für die Spracherzeugung wird das verlustbehaftete Rohrmodell verwendet (Bild 1). Das Modell basiert auf dem unverzweigten Kreuzgliedkettenfilter, erweitert um die Modellierung der frequenzabhängigen Vokaltraktverluste

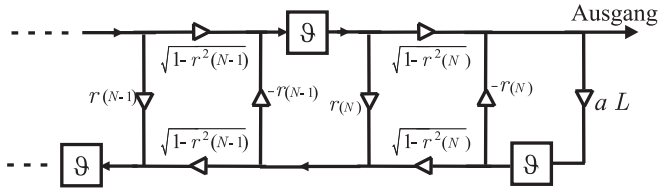

Bild 1: Verlustbehaftetes Rohrmodell mit den Verlustsystemen $\alpha L(z)$ und $\vartheta(z)$ sowie den Reflexionskoeffizienten $r(n)$.

[1]. Der konzentrierte Verlust infolge der Lippenabstrahlung wird durch durch das System $\alpha L(z)$ am Systemausgang modelliert. Die verteilten Vokaltraktverluste infolge viskoser Reibung, Wärmeleitung und Wandvibrationen werden durch die Verlustsysteme $\vartheta(z)$ berücksichtigt. Die Parameter von ϑ sind nach Literaturangaben optimiert. Die freien Parameter des Rohrmodells sind die Reflexionskoeffizienten, die aus dem Sprachsignal *s* und der Systemfunktion $H(\omega)$ des Rohrmodells durch Minimierung des spektralen Abstandsmaßes

$$e = \frac{1}{\pi}\int_0^{\pi} \left|\frac{S'(\omega)}{H(\omega)}\right|^2 \mathrm{d}\omega \quad \rightarrow \quad \min \qquad (1)$$

geschätzt werden. Der Fehler *e* ist aus der inversen Filterung abgeleitet und wird mittels eines Gradientenverfahrens minimiert. Um den Einfluß der Anregung und Abstrahlung in dem Sprachsignal zu beseitigen, wird das Sprachsignal *s* zuvor mit einer wiederholten adaptiven Präemphase gefiltert. Das gefilterte Sprachsignal s' wird dann in überlappende Segmente aufgeteilt, aus denen die Reflexionskoeffizienten geschätzt werden [2]. Die Schätzung und Filterung der Präemphase wird für stimmhafte Abschnitte und Signalabschnitte von stimmlosen Frikativen diphonweise vollzogen. Als Resultat erhält man abschnittsweise ähnlich der LPC-Codierung einen Modellparametervektor $\boldsymbol{r}_i$ für jedes Segment *i*. Für die Synthese der Diphone werden die geschätzten Modellparametervektoren sukzessive für das Rohrmodell verwendet. Lautverlängerungen oder Verkürzungen können durch Verdopplung oder Auslassen von Parametervektoren realisiert werden. Als stimmhafte Anregung für das Rohmodell wird eine Impulsfolge verwendet, die durch das System von reellen Polstellen der Präemphasekoeffizienten gefiltert wird und mit hochpaßgefiltertem Rauschen additiv überlagert wird. Bei stimmhaften Frikativen hingegen wird durch gefiltertes Rauschen angeregt.

Analyse der Diphone

Durch die Analyse der Diphondatenbank mit Hilfe des Rohrmodells werden die Diphone durch Sätze von Modellparametervektoren beschrieben. Ausnahmen bilden hierbei die stark instationären stimmlosen Lautabschnitte, wie z.B. die der Explosive, welche durch ihr Zeitsignal repräsentiert werden. Bei der Diphonsynthese werden die Sätze von Parametervektoren der Diphone verbunden. Diese modellbasierte Diphonverkettung weist zunächst auch wie andere Konkatenationsverfahren Unstetigkeiten an den Verkettungsstellen auf, da sich die Diphone nicht mehr in ihrer ursprünglichen Umgebung befinden [3]. Die Unstetigkeiten sind für verschiedene Diphonkombinationen unterschiedlich stark ausgeprägt. Die Verkettungsstellen zwischen den Diphonen werden durch einen linearen Übergang in der Parameterdarstellung der logarithmierten Flächen geglättet [2]. Diese Vorgehensweise führt allerdings nicht bei allen Kombinationen der Diphone zum gewünschten Erfolg, da immer noch störende Artefakte auftreten können. Daher ist es günstig, wenn sich die Parametersätze der Diphongrenzen gleicher Laute schon vor dem Verkettungsalgorithmus möglichst wenig unterscheiden. Diese Angleichung wird durch eine zweistufige Analyse der Diphone erzielt: In der ersten Stufe werden die Koeffizienten der Diphonsegmente durch Minimierung des Fehlermaßes (1) geschätzt, wie in [2] beschrieben. Dabei startet der Optimierungsalgorithmus von einer neutralen Startkonfiguration in der alle Reflexionskoeffizienten gleich Null sind. Die Ergebnisse der

ersten Analyse werden verwendet, um repräsentative Parametersätze $\boldsymbol{r}^{\theta}$ für die Phoneme θ zu erhalten. Dafür werden die Paramtersätze der Diphongrenzen gleicher Phoneme θ gemittelt. Die Mittelung der Modellparameter wird in der Darstellung der LAR (Log Area Ratio) durchgeführt; der resultierende Parametersatz wird anschließend wieder in die Darstellung der logarithmierten Flächen überführt. Die gemittelten Parametervektoren $\boldsymbol{r}^{\theta}$ stellen repräsentative Vokaltraktkonfigurationen der Phoneme θ dar und werden als Startvektoren des Optimierungsverfahrens in einer erneuten Analyse der Diphone verwendet; dies stellt die zweite Stufe der Analyse dar. Dabei werden die Startkonfigurationen mit einem Faktor gewichtet, der an den Diphongrenzen den Wert Eins besitzt, und an den Stellen zwischen den beiden Lauten innerhalb des Diphons Null ist; zwischen diesen beiden Positionen wird der Faktor linear interpoliert. Dadurch ist ein von Segment zu Segment stetiger Verlauf der Startkonfiguration gewährleistet. Tabelle 1 zeigt die Varianzen der Parametervektoren von Diphongrenzen gleicher Phoneme θ, ermittelt aus der Diphondatenbank de1 [4]. Die Varianzen $\bar{\sigma}^2_{\theta 1}$ und $\bar{\sigma}^2_{\theta 2}$ der ersten bzw. zweiten Analysestufe sind in der Darstellung der LAR berechnet. Es ist zu sehen, daß die Variationen der Koeffizientensätze durch die weitere Analyse verringert sind. Dies lässt sich dadurch erklären, daß durch die gewählten Startkonfigurationen der zweiten Stufe die Parametervektoren weniger weit auseinander konvergieren. Damit stehen für die Diphonverkettung Parametersätze zur Verfügung, die im Vorhinein weniger Unstetigkeiten aufweisen.

θ	$\bar{\sigma}^2_{\theta 1}$	$\bar{\sigma}^2_{\theta 2}$	θ	$\bar{\sigma}^2_{\theta 1}$	$\bar{\sigma}^2_{\theta 2}$
/a/	0.22	0.12	/j/	0.47	0.33
/a:/	0.19	0.10	/n/	0.45	0.30
/I/	0.42	0.24	/v/	0.42	0.24
/i:/	0.50	0.18	/z/	0.54	0.26
/@/	0.34	0.18	/l/	0.47	0.25
/o:/	0.41	0.15	/f/	0.18	0.12
/u:/	0.48	0.22	/S/	0.33	0.11

Tabelle 1: Varianzen der LAR-Parameter der Diphongrenzen nach der ersten Stufe $\bar{\sigma}^2_{\theta 1}$ und nach der zweiten Stufe $\bar{\sigma}^2_{\theta 2}$ der Analyse (repräsentative Auswahl).

Bild 2 zeigt ein Beispiel einer Verkettung der Diphone [z-o:] und [o:-p]. An den Betragsgängen des Rohrmodells kann die Bewegung der Formanten im Übergang zwischen zwei Diphonen festgestellt werden. Es ist zu sehen, daß sich die ersten beiden Formanten mit den Parametersätzen der ersten Analysestufe kreuzen (vgl. Bild 2(a)) im Gegensatz zu den Ergebnissen mit der zweiten Analysestufe (vgl. Bild 2(b)). Die Kreuzungen der Formanten in Bild 2(a) treten im mittleren stationären Teil des Vokals auf und können daher als unrealistisch angesehen werden. Dieser Fehlverlauf der Formanten tritt hier durch die zweite Analyse nicht auf, was darin begründet werden kann, daß die Koeffizientensätze an den Verkettungsstellen durch den linearen Parameterübergang geringer modifiziert werden. Neben diesen Fehlverläufen der Formanten können starke Änderungen der Koeffizienten im Diphonübergang störende Artefakte im Zeitsignal bedingen, wobei auch hier die Verwendung der zweiten Analysestufe sich als günstiger erweist.

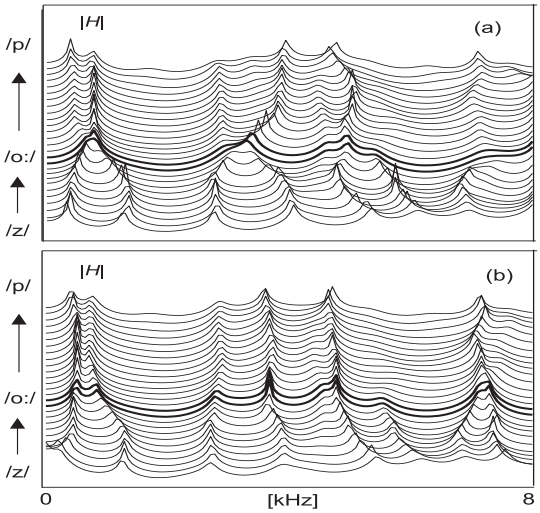

Bild 2: Betragsgänge des Rorhmodells an der Verkettungsstelle der Diphone [z-o:] und [o:-p]: (a) Ergebnisse mit Analysestufe 1, (b) mit Analysestufe 2. Die Betragsgänge der Diphongrenzen sind hervorgehoben.

Zusammenfassung

In diesem Beitrag konnte gezeigt werden, wie die Verkettungsergebnisse von Diphonen mit dem verlustbehafteten Rohrmodells verbessert werden. Dies wird durch eine zweistufige Analyse der Diphone ermöglichst, welche die Parametersätze an den Verkettungsstellen besser angleicht. Synthetisierte Beispiele zeigen, daß durch die Angleichung der Parametersätze an den Diphongrenzen eine Verbesserung der Sprachqualität erzielt wird.

Literatur

[1] Schnell K., Lacroix A.: "Analysis of lossy vocal tract models for speech production", Proc. EUROSPEECH-2003, Geneva Switzerland pp. 2369-2372, 2003.

[2] Schnell K., Lacroix A.: „Verkettung von Spracheinheiten für die Spracherzeugung mittels verlustbehafteter Rohrmodelle", Studientexte zur Sprachkommunikation: Band 30, 15. Konferenz ESSV-2004, Cottbus, S. 163-170. 2004.

[3] D.T. Chappell and J.H.L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis", Speech Communication (36) pp. 343-374, 2002.

[4] Englert, F.: "Acquisition of a Diphone Database for German" In: Wodarz, H.-W. (Ed.): Forum Phoneticum 63, Speech Processing, Hector-Verlag Frankfurt am Main, 1997, pp. 23–32.