Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

SEMINARARBEIT

Neuronale Netze in der statistischen parametrischen Sprachsynthese: Ein Blick auf Natürlichkeit und Performanz

vorgelegt von

Jan Lütjen

MIN-Fakultät

Fachbereich Informatik

Studiengang: Software-System-Entwicklung

Matrikelnummer: 6295286

Inhaltsverzeichnis

1 Einleitung

Im Rahmen des Proseminars „Verarbeitung gesprochener Sprache“ im Sommersemester 2020 durfte ich mich näher mit der statistischen parametrisierten Sprachsynthese(SPSS) auf Basis von Hidden Markov Modellen(HMMs) beschäftigen.

Im Vergleich zu konkatenativen Vorgehensweisen bei der Synthese bietet die SPSS einige Vorteile, u.a. die Möglichkeit Stimmcharakteristika flexibel anzupassen und der vergleichsweise niedrige Speicherplatzbedarf, die ihr in den letzten Jahren zu steigender Popularität verholfen haben [1]. Trotzdem bleibt sie bisher häufig qualitativ hinter der sprachlichen Natürlichkeit anderer Synthesemethoden zurück [2, 3]. Vor diesem Hintergrund gewinnen andere Lösungsansätze innerhalb der SPSS zusehends an Bedeutung. Einer dieser Ansätze ist der Einsatz neuronaler Netze an Stelle der HMMs [4]. Hieran anknüpfend beschäftigt die Seminararbeit sich mit der Frage, welche Unzulänglichkeit(en) der HMM-basierten Sprachsynthese durch die Nutzung neuronaler Netzwerke umgangen werden sollen und welche Performanz und Natürlichkeit bei der Synthese diese neuen Lösungsansätze im Vergleich aufweisen. Dabei konzentriert sich die Arbeit primär auf den Teilbereich der Generierung der für die Sprachsynthese benötigten akustischen Parameter. Unterschiede, die sich durch den Trainingsaspekt der sprachlichen Modelle ergeben, werden nicht beleuchtet.

Die Arbeit gliedert sich im Folgenden in 3 Abschnitte. Im ersten Abschnitt wird die HMM-basierte Sprachsynthese kurz vorgestellt und näher auf ihre Unzulänglichkeiten eingegangen, die durch den Einsatz von neuronalen Netzen umgangen werden sollen. Mit diesen Netzen beschäftigt sich anschließend der zweite Abschnitt, der einen Überblick über für die Sprachsynthese relevante neuronale Netztypen liefert. Der konkrete Einsatz neuronaler Netze in der SPSS wird anhand von Forschungsbeiträgen und Anwendungsbeispielen vorgestellt. Die Unterschiede in Performanz und Natürlichkeit zu der HMM-basierten Synthese werden dann in Abschnitt drei zusammengetragen und ausgewertet.

2 Hidden Markov Modell-basierte Sprachsynthese

Der Sprechvorgang kann durch einen digitalen Filter, einen sogenannten Vocoder, nachgebildet werden [5, Sec. II.A.]. Im Vocoder werden aus Erregungssignalen durch einen Resonanzfilter Sprachschallwellen modelliert. Für diesen Vorgang benötigt der Vocoder Informationsparameter über Stimmhaftigkeit, Grundfrequenz und Spektraleigenschaften der zu synthetisierenden Schallwellen. Aus einer Folge solcher Parametermengen lässt sich durch den Vocoder eine Sprachwellenform bilden. Daher ist eine der zentralen Problemstellungen der Sprachsynthese die Generierung entsprechender akustischer Parameter aus linguistischen Spezifikationen, wie z.B Phonemfolgen und -verweildauer eines zu synthetisierenden Textes. Bei der statistischen parametrisierten Sprachsynthese wird dieser Abbildungsprozess mit Hilfe von statistischen Modellen erreicht. Schwerpunkt der Forschung liegt dabei auf dem Hidden Markov Modell [6, p. 447].

2.1 Hidden Markov Modell

Das HMM beschreibt zwei gekoppelte Zufallsprozesse [7, Ch. 5]. Ein Markov-Prozess, bestehend aus einer Menge an verdeckten Zuständen mit jeweiligen Übergangswahrscheinlichkeiten, bildet den ersten Zufallsprozess. Abhängig von den Zuständen zugeordneten multivariaten Wahrscheinlichkeitsverteilungen erzeugt der zweite Zufallsprozess für jeden diskreten Zeitpunkt eine Beobachtung. Das Durchlaufen einer Zustandssequenz erzeugt also eine Folge von Beobachtungen.

2.2 Sprachparameter-Generierung durch HMMs

In der SPSS beschreiben HMMs einzelne Phoneme. Für jedes Phonem in einer vorliegenden Phonemfolge werden die wahrscheinlichste Zustands- und Beobachtungsfolge seines HMM ermittelt. Um zu vermeiden, dass die einzelnen Zustände stets ihre durchschnittlichen Beobachtungen und so Sprünge an Zustandsgrenzen erzeugen, werden auch dynamische Eigenschaften natürlicher Sprache berücksichtigt [5, Sec. C.2)]. Hierfür wird angenommen, dass jede Beobachtung sich aus einem statischen und einem dynamischen Teil zusammensetzt. Dieser dynamische Teil wird aus den statischen Teilen benachbarter Beobachtungen errechnet und schränkt die Lösungsmenge generierbarer Beobachtungen zu einem Zeitpunkt der Zustandsfolge ein. Auf diese Weise können auch zeitliche Abhängigkeiten durch den HMM-Ansatz abgebildet werden.

2.3 Limitierende Faktoren der HMM-basierten Sprachsynthese

Gegenüber der anderen Synthese-Technik der „third generation“ [6, p 447], der Unit-Selection, bietet die HMM-basierte Synthese viele Vorteile. Nennenswert sind hierbei die Möglichkeit Stimmcharakteristika flexibel anzupassen und der vergleichsweise niedrige Speicherplatzbedarf [4]. Wie aber Black *et al*. [2] und diverse Ergebnisse der Blizzard Challenge über die letzten Jahre gezeigt haben [3], reicht die SPSS auf HMM-Basis nicht an die Natürlichkeit anderer Synthesemethoden heran. Dies ist auf die Qualität des eingesetzten Vocoders, die Genauigkeit des Sprachmodells und einen überhöhten „Weichzeichner-Effekt“ durch Berücksichtigung der dynamischen Spracheigenschaften zurückzuführen.

3 Sprachsynthese auf Basis neuronaler Netze

Bei der Sprachsynthese auf Basis neuronaler Netze(NNs) kommen bei der Abbildung sprachlicher Spezifikationen auf akustische Merkmale NNs zum Einsatz. Ein NN besteht aus einzelnen Knoten, den Neuronen, die über gerichtete Kanten miteinander verbunden sind. Ein Neuron kann Signale anderer Neuronen als Eingaben empfangen. Diese werden gewichtet an eine Aktivierungsfunktion übergeben, die den Ausgabewert des Neurons unter Berücksichtigung eines Schwellenwertes berechnet. Die einzelnen Neuronen sind in Schichten angeordnet. Jedes Netz besitzt eine Ein- und eine Ausgabeschicht. Beim Einsatz von NNs in der SPSS werden über die Eingabeschicht die linguistischen Spezifikationen in das Netz eingespeist. Die Ausgabeschicht stellt die entsprechenden akustischen Parameter zur Verfügung.

3.1 Netztypen

Der Typ eines Netzes wird durch seine Topologie bestimmt.

3.1.1 Deep Neural Networks

Deep Neural Networks(DNNs) beschreiben sogenannte Feed-Forward-Netze(FFNNs) mit mehreren Schichten. In FFNNs sind Neuronen nur mit Neuronen der nächsthöheren Schicht verbunden.

3.1.2 Recurrent Neural Networks

Recurrent Neural Networks(RNNs) besitzen anders als FFNNs auch rückgerichtete Kanten. Diese rekurrenten Kanten senden meist mit einer zeitlichen Verzögerung. So können bei einer schrittweisen Verarbeitung Neuronenausgaben des vergangenen Schrittes wieder als Eingabe verarbeitet werden. Durch diese Rückkopplungen wird ein dynamisches Verhalten des Netzes ermöglicht.

Eine Sonderform der des RNN stellt das long short-term memory RNN(LSTM-RNN) dar [8]. Durch sogenannte „memory blocks“ können Informationen in den Gedächtniszellen des Netzes deutlich länger gehalten werden als bei herkömmlichen RNNs.

3.2 Forschungs- und Anwendungsübersicht

Mittlerweile gibt es eine Vielzahl an Forschungsbeiträgen, die sich mit dem Einsatz von NNs in der Verarbeitung gesprochener Sprache beschäftigen. Im Folgenden werden drei wissenschaftliche Beiträge vorgestellt, in denen NNs die Rolle des generativen Modells innerhalb der SPSS übernehmen.

3.2.1 Überblick über NNs als Modellalternative zu HMMs

In seiner Arbeit von 2015 gibt Heiga Zen einen Überblick über die in der SPSS eingesetzten generativen Modelle [4]. Dabei geht er insbesondere auch auf die Nutzung von NNs ein.

So erreichen DNNs durch Testhörer im Bereich Natürlichkeit eine höhere Bewertungspunktzahl als HMMs. Allerdings benötigen sie mehr Zeit für die Synthese. Es werden nämlich nicht wie beim HMM-Ansatz Entscheidungsbäume für jeden Zustand durchlaufen, sondern es müssen für jeden diskreten Zeitpunkt Matrizenmultiplikationen durchgeführt werden, was deutlich mehr Rechenaufwand erfordert.

Für RNNs bzw. LSTM-RNNs siedelt sich die für die Synthese benötigte Zeit zwischen HMM und DNN an. Bei der subjektiven Bewertung der Natürlichkeit schneiden das LSTM-RNN sogar noch besser als das DNN ab.

3.2.2 DNN statt HMM

Zen, Senior and Schuster vergleichen in ihrer Arbeit von 2013 SPSS auf Basis von HMMs und DNNs [1]. Hierzu wurden 2 entsprechende Synthese-Systeme am selben Sprachkorpus trainiert und die Sprachergebnisse anschließend einer objektiven Bewertung anhand eines Kriterienkataloges aus Spektral- und Erregungsparametern und einer subjektiven Bewertung durch Testhörer unterzogen.

Bei der objektiven Bewertung schneidet die DNN-basierte Synthese hier bei 3 von 4 geprüften Kriterien besser ab. Nur bei der Bestimmung der logarithmierten Grundfrequenz erreicht die Synthese auf HMM-Basis eine höhere Genauigkeit.

Die Bewertung der Natürlichkeit der DNN-Synthese durch die Testhörer ist signifikant besser als die des HMM.

3.2.3 LSTM-RNN-basierte Sprachsynthese auf Mobilgeräten

Zen et al. beschreiben in ihrer Arbeit aus 2016 Optimierungsmöglichkeiten eines LSTM-RNN-basierten Sprachsynthese-Systems für Mobilgeräte [9]. Im Rahmen der Arbeit werden u.a. die Leistungsdaten dieses Systems mit denen eines HMM-basierten verglichen.

Das optimierte LSTM-RNN schneidet sowohl bei Schnelligkeit und Latenz der Synthese als auch bei der durch Testhörer bewerteten Natürlichkeit besser ab. Der Leistungstest wurde auch für Smartphones mit einem älteren CPU-Modell ohne modernes NEON Befehls-Set durchgeführt. Hier war das LSTM-RNN 15-22% langsamer, wies aber immer noch eine niedrigere Latenz auf.

4 Auswertung und Beurteilung der SPSS auf Basis von NNs

Die Kriterien, anhand derer die vorliegenden Forschungsarbeiten die NN-basierte Sprachsynthese beurteilen und die für die Forschungsfrage dieser Arbeit relevant sind, lassen sich in Natürlichkeit und Performanzkriterien aufteilen.

4.1 Natürlichkeit

In allen drei Arbeiten werden die durch die NN-basierte Sprachsynthese erzeugten Sprachbeispiele von den Testhörern als natürlicher empfunden. Einschränkend muss erwähnt werden, dass [4] bei der Einordnung des DNN die wissenschaftliche Arbeit von Zen, Senior und Schuster aus 3.2.2 referenziert. Aus den Vergleichswerten aus Spektral- und Erregungsparametern aus [1] lassen sich zwar Rückschlüsse auf die Genauigkeit der DNN-basierten SPSS ziehen, aber auf Grund mangelnder Korrelation keine Aussagen über ihre Natürlichkeit treffen.

4.2 Performanz

Im Gegensatz zur Natürlichkeit bietet sich bei der Performanz von NNs in der SPSS kein so eindeutiges Bild. Während in [4] HMMs eine schnellere Synthesezeit aufweisen als die angebotenen NNs, sehen Zen et al. [9] das LSTM-RNN vor der HMM-basierten Synthese.

Latenzzeiten sind je nach angewendetem NN gleich bis geringer als bei HMMs.

Zen, Senior und Schuster treffen in [1] keine Aussagen über Schnelligkeit und Latenz. Allerdings weist das DNN-Modell bei ihnen eine höhere Genauigkeit bei der Vorhersage von Sprachparametern auf.

5 Fazit

Diese Arbeit hat das Abschneiden in Performanz und Natürlichkeit auf NNs basierender Sprachsynthese in aktuellen Forschungsarbeiten untersucht. Dabei wurde das Potential dieser Herangehensweise bei der Synthese deutlich, die Schwächen der HMM-basierten SPSS bei der Natürlichkeit der produzierten Sprachschallwellen zu umgehen. Die Performanz betreffend, gestaltet sich die Beurteilung nicht ganz so einfach. Hier bleiben die NNs auf Grund des hohen Berechnungsaufwands bei der Synthese hinter den HMMs zurück. Zen et al. [9]haben jedoch gezeigt, dass diese Lücke sich mit stetig steigender Rechenleistung und Optimierung der Netztopologien bald schließen wird.

6 Literaturverzeichnis

[1] H. Zen, A. Senior, and M. Schuster, „Statistical parametric speech synthesis using deep neural networks,“ in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP'13*, Vancouver, British Columbia, Canada, 2013, pp. 7962-7966.

[2] A. Black, H. Zen, and K. Tokuda, “Statistical Parametric Speech Synthesis,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP'*07, Honolulu, Hawaii, USA, 2007, Vol. 4, pp. 1229-1232.

[3] T. Merritt, J. Yamagishi, Z. Wu, O. Watts, and S. King, „Deep Neural Network Context Embeddings for Model Selection in Rich-Context HMM Synthesis,“ in *Proceedings of Interspeech 2015*, Dresden, Germany, 2015.

[4] H. Zen, “Acoustic Modeling in Statistical Parametric Speech Synthesis - from HMM to LSTM-RNN,” in *Proc. MLSLP*, 2015.

[5] K. Tokuda, et al, "Speech Synthesis Based on Hidden Markov Models," *Proceedings of the IEEE,* Vol. 101, No. 5, pp. 1234-1252, 2013.

[6] P. Taylor, *Text-to-Speech Synthesis.* Cambridge: Cambridge university press, 2009.

[7] B. Pfister, and T. Kaufmann, *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Zürich: Springer-Verlag, 2008.

[8] S. Hochreiter, and J. Schmidhuber, „Long Short-Term Memory,“ *Neural Computation 9*, pp. 1735-1780, 1997.

[9] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, “Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices,” in *Proc. Interspeech*, 2016.