Nutzbarkeit von modellierten Phonemfolgen zur Erkennung von unbekannten Wörtern in phonembasierten Spracherkennern

Matthias Deppermann gen. Esser[1], Jan Wellmann[1], Niko Moritz[1], Stefan Goetze[1]

[1] *Fraunhofer Institut für digitale Medientechnologie IDMT, Projektgruppe Hör-Sprach- und Audiotechnik, 26129 Oldenburg, E-Mail: matthias.esser@idmt.fraunhofer.de*

1. Einleitung

Der Wortschatz eines Schlüsselworterkenners kann durch das Hinzufügen neuer Phonemsequenzen erweitert werden. Häufig werden dafür Triphon-Modelle benötigt, die nicht im Trainingsmaterial enthalten sind.

Bei zwei häufig genutzten Verfahren werden fehlende Triphone durch entsprechende Monophone ersetzt oder anhand einer datengetriebenen Ähnlichkeitsbeziehung der Hidden-Markov-Modelle (HMM) durch vorhandene Modelle abgebildet. Über Entscheidungsbäume mit Fragen zu dem linken und rechten Phonemnachbarn wird so ein ähnliches Triphon gefunden, welches das fehlende Triphon abbilden kann.

In dieser Arbeit wird eine weitere Methode untersucht, in der fehlende Triphone synthetisch generiert werden, in dem das gesuchte akustische Modell (AM) aus zwei ähnlichen und vorhandenen Triphon-Modellen zusammengesetzt wird.

Die Leistung dieser drei Verfahren wird anhand unbekannter Äußerungen und Phantasiewörter mit einem phonembasierten Spracherkenner verifiziert.

2. Grundlagen

Der Klang eines Phonems hängt im Allgemeinen von dem vorangegangenen und dem nachfolgenden Phonem ab, da diese mehr oder weniger stark durch ihre Nachbarlaute (Kontexte) geprägt werden (Koartikulation). Ein Triphon stellt den Laut des mittleren Phonems der Folge dar. Durch das Berücksichtigen von verschiedenen Koartikulationen, wird die Worterkennerrate (WR) verbessert.

Es wird zwischen einem physikalischen und logischen Triphon unterschieden. Ein physikalisches Triphon ist in den Trainingsdaten enthalten und besitzt ein eigenes AM. Das logische Triphon ist dagegen möglicherweise nicht im Trainingsmaterial enthalten und besitzt daher auch kein eigenes AM.

2.1 Zustandsgleichstellung mit einem Monophon

Eine Möglichkeit ist, ein fehlendes Triphon durch das im AM enthaltene Monophon zu ersetzten. Bei dieser Methode wird der Kontext des jeweiligen Kernphonems vernachlässigt. Ein Beispiel ist in Abbildung 1 gezeigt.

b – a + r → [a]

Abbildung 1: Triphon wird mit Monophon gleichgestellt.

2.2 Zustandsgleichstellung mit Hilfe von phonetischen Entscheidungsbäumen

Das fehlende logische Triphon wird mit einem ähnlichen physikalischen Triphon gleichgestellt und nimmt damit das AM eines bekannten Triphon an. In Abbildung 2 ist ein solches Beispiel gegeben.

b – a + r → p – a + s

Abbildung 2: Triphon wird mit ähnlichem Triphon gleichgestellt.

Die in Abbildung 3 dargestellten Entscheidungsbäume helfen beim Finden ähnlicher Triphone. Ausgehend von der Wurzel des binären Entscheidungsbaumes werden Ja-Nein-Fragen zu dem linken und rechten Nachbarn gestellt, um alle möglichen Triphone zu gruppieren [1]. Anschließend wird die Likelihood der Triphon-Gruppe bestimmt, welcher durch jede weitere Aufspaltung zunimmt. Dieser Prozess wird fortgeführt bis das Ende des Entscheidungsbaumes erreicht ist oder die Likelihood Steigerung unterhalb einer gewählten Schwelle fällt. Die HMM Zustände der so gefundenen Triphon-Gruppe werden dann verbunden. Diese Methode entspricht dem Standard in aktuellen Spracherkennersystemen.

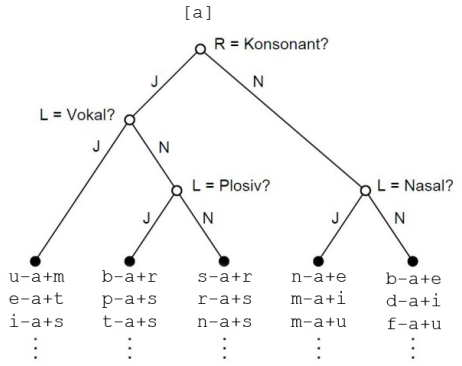

Abbildung 3: Phonetischer Entscheidungsbaum (Quelle [2])

3. Hauptteil

3.1 Generierung eines synthetischen Triphons

In dieser Arbeit wird eine weitere Methode untersucht, in der ein fehlendes Triphon aus zwei physikalisch vorhandenen Triphonen generiert wird. Dazu werden zunächst zwei vorhandene Triphone mit gleichem zentralen Laut ausgewählt, die zum gesuchten Triphon einen ähnlichen linken und rechtem Kontext aufweisen. Die Ähnlichkeit der Nachbarn wird anhand der phonetischen Entscheidungsbäume bestimmt.

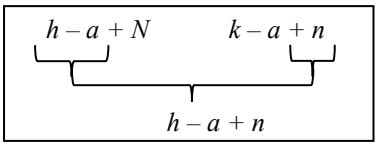

Abbildung 4: Generieren des Triphones „h – a + n“

In Abbildung 4 ist ein Beispiel gezeigt. Das gesuchte Triphon „h-a+n“ wird hierbei aus den ersten beiden HMM Zuständen von „h-a+N“ und aus dem dritten Zustand des physikalischen Modells „k-a+n“ zusammengesetzt. Vorausgesetzt ist hierbei, dass jedes Triphon aus drei emittierenden HMM Zuständen besteht.

```
~h "h-a+n"
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
~s "ST_a_2_21"
<STATE> 3
~s "ST_a_3_28"
<STATE> 4
~s "ST_a_4_22"
~t "T_a"
<ENDHMM>
```

Abbildung 5: Auszug aus dem HMM

In Abbildung 5 ist ein Ausschnitt aus dem generierten HMM mit den zugehörigen Zuständen zu sehen. Die Zustände werden so mit dem passende AM verknüpft.

4. Evaluierung

4.1 Testumgebung zur Evaluierung

Zur Evaluierung ist die Spracherkennungs-Engine „Julius“ auf einem eingebetteten System verwendet worden. Das in dieser Arbeit verwendete HMM ist aus 26448 Tonaufnahmen mit 1638 verschiedenen Äußerungen erstellt. Die meisten Äußerungen stammen aus Trainingsmaterial für die Sprachsteuerung eines assistiven Roboters, eines Fernsehers sowie Kommandos zur sprachgesteuerten Bedienung eines Terminkalenders.

4.2 Ergebnisse

Im Diagramm 1 sind unbekannte Wörter gegeben, welche Triphone enthalten, die nicht im Trainingsmaterial vorhanden sind. Zu sehen sind die WR´s der drei getesteten Verfahren.

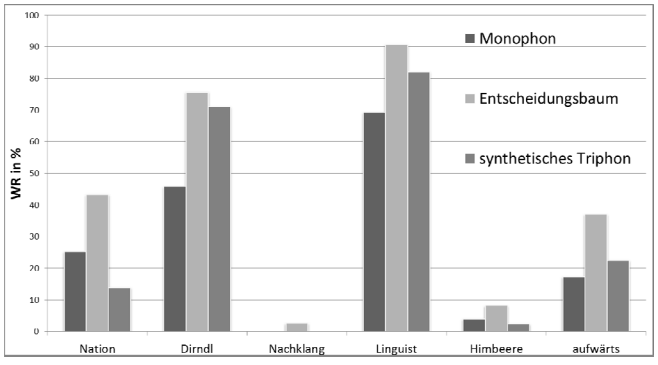

Diagramm 1: Worterkennerrate bei unbekannten Wörtern

Anhand von Phantasiewörtern wird der Trainingserfolg bei Dialekten bzw. absichtlicher Falschaussprache realisiert.

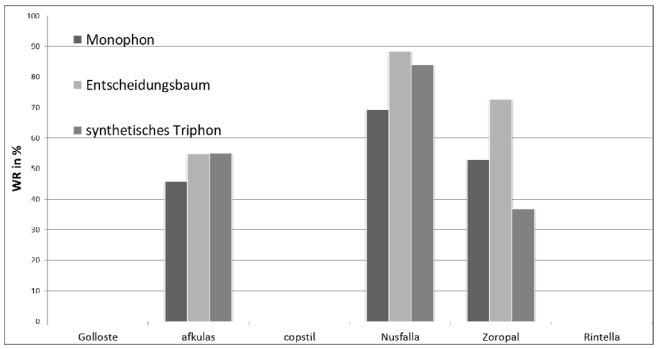

Diagramm 2: Worterkennerrate bei Phantasiewörtern

Bei der in Diagramm 2 zu sehenden Auswahl an Phantasiewörtern, werden nur drei von sechs Äußerungen erkannt. Auch durch das Generieren synthetischer Triphone kann die WR im Vergleich zu der Methode mit den Entscheidungsbäumen nicht verbessert werden.

Die Zustandsgleichsetzung mit Hilfe von phonetischen Entscheidungsbäumen hat im Durchschnitt eine um 10,89% bessere WR, im Vergleich zum Generieren eines synthetischen Triphones.

5. Zusammenfassung

Alle drei vorgestellten Verfahren ermöglichen die Detektion unbekannter und untrainierter Wörter, jedoch muss bei der Verwendung von synthetischen Triphonen im Vergleich zu den beiden weiteren Methoden eine deutlich verringerte Erkennerleistung in Kauf genommen werden. Am besten hat von allen untersuchten Varianten die in der Sprach-erkennung am häufigsten verwendete Methode funktioniert, welche phonetische Entscheidungsbäume und Ähnlichkeits-beziehungen der AM im Trainingsprozess ausnutzt. Daraus lässt sich schließen, dass die Anfangs- und Endzustände in einem HMM nicht alleinig dem Phonem-Kontext zugeordnet werden können, sondern dass dieser über mehrere HMM Zustände abgebildet wird. Daher ist ein auseinander Pflücken und neu Zusammensetzen von Zuständen aus verschiedenen HMMs nicht ohne weiteres praktikabel. Die verketteten Zustände, des durch Modellierung generierten HMM, bilden den phonetischen Zusammenhang somit schlechter ab als ein ähnliches Triphon.

6. Literatur

[1] Pfister, B.; Kaufmann, T .: Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung. Springer, Berlin, Heidelberg, 2008

[2] Euler, S.: Grundkurs Spracherkennung. Vom Sprachsignal zum Dialog ; Grundlagen und Anwendung verstehen ; mit praktischen Übungen. Vieweg, Wiesbaden, 2006