State-Tying – Wie sich Nachbarphoneme auf das Trainieren von Spracherkennern auswirken

von Lukas Seifert

Proseminararbeit in Verarbeitung gesprochener Sprache

Abstract

Zum Trainieren eines HMM-basierten Spracherkenners mit großem Vokabular bieten sich anfangs Phoneme an, jedoch berücksichtigen diese keine Koartikulationseffekte. Deswegen eignen sich Triphone, welche das linke und rechte Phonem als Kontext der Aussprache mitberücksichtigen.
Phoneme gibt es in jeder Sprachen etwa zwischen 10 und 80, Triphone hingegen deutlich mehr, dementsprechend liegen aufgrund eines begrenzten Trainingskorpus weniger Daten zum Schätzen der Parameter jedes Triphons vor.
Deswegen soll die Anzahl der Parameter eingeschränkt werden, indem ein sogenanntes State-Tying angewandt wird, bei welchem mehrere Zustände der Triphone mit akustischen Ähnlichkeiten zu einem verknüpft werden.
Es gibt verschiedene Verfahren zum State-Tying, die im Allgemeinen als Bottom-Up und Top-Down-Verfahren bezeichnet werden.
Das wird in dieser Arbeit näher betrachtet werden und beide Verfahren verglichen, sowie der Nutzen des State-Tyings in Bezug auf eine Verbesserung der Wortfehlerrate untersucht.

Inhaltsverzeichnis

1. Einführung

In der Spracherkennung gibt es verschiedene Ansätze, oft werden statistische Verfahren eingesetzt, die meist auf Hidden Markov Modelle (HMM) basieren. Sie bestimmen an einer Sprachäußerung die wahrscheinlichste Sequenz an Spracheinheiten und geben somit das am wahrscheinlichsten Gesagte aus.
Ein HMM besitzt verschiedene Parameter und Wahrscheinlichkeiten, welche als Gaußverteilung dargestellt sind. Abhängig vom Modell können dies Normalverteilungen oder Mischverteilungen sein.
Die Wahrscheinlichkeiten werden anfangs nur geschätzt und durch Training optimiert, sodass eine möglichst niedrige Wortfehlerrate (WER) vorliegt. Meistens wird der Baum-Welch-Algorithmus angewendet, durch den die Parameter an die Trainingsdaten angepasst werden.
Die Spracheinheiten, die Wörter, Phoneme, Biphone, Triphone, etc. sein können, bilden die Grundelemente der HMM. Alle Grundelemente der HMM besitzen Vor- und Nachteile; je komplexer das Modell, desto mehr Trainingsdaten werden benötigt. [Young & Woodland 93]

Wörter sind in kontinuierlich gesprochener Sprache mit großem Vokabular unpraktisch und werden kaum verwendet, da eine zu große Menge an Trainingsdaten notwendig wäre, welche mit der Anzahl der verschiedenen Elemente steigt. Für jedes Wort müssten genügend Daten vorliegen, damit der Spracherkenner sie später erkennen kann.
Phoneme sind in dieser Hinsicht praktisch, da es von ihnen nur jeweils zwischen 10 und 80 in den verschiedenen Sprachen gibt und sie somit ausreichend Trainingsdaten für jedes Phonem zur Verfügung stehen. Jedoch sinkt mit ihnen die Genauigkeit, da sie keine Koartikulationseffekte berücksichtigen; der Einfluss der Nachbarphoneme auf die Aussprache, sie werden deswegen auch als kontextunabhängig bezeichnet.
Biphone und Triphone sind kontextabhängig, sie berücksichtigen ihre Nachbarphoneme als Kontext der Aussprache. Ein Triphon besteht aus einem Kernphonem und seinem linken und rechten Kontextphonem, geschrieben als a|n|r, wobei n das betrachtete Kernphonem ist und a und r die Aussprache des n beeinflussen.
In der Praxis werden sowohl wortinterne als auch wortübergreifende Triphone verwendet werden, da in fließender Sprache das Ende eines Wortes Einfluss auf die Aussprache des Anfangs des nächsten Wortes hat. Triphone haben eine hohe Genauigkeit, dafür steigt die Anzahl der Parameter aufgrund der vielen Kombinationsmöglichkeiten, das hat wiederum Einfluss auf die zur Verfügung stehenden Trainingsdaten für jedes Triphon; viele der Triphone kommen nur selten oder gar nicht vor, für sie gibt es dementsprechend kaum bis gar keine Daten. Das Ziel ist es, ein gutes Mittelmaß zwischen Genauigkeit und zur Verfügung stehenden Trainingsdaten pro Parameter zu finden. In der Spracherkennung wird deswegen häufig eine Mischung aus Monophonen, Biphonen und Triphonen angewandt, mit der Methode des State-Tyings hingegen ist es möglich, ausschließlich Triphone zu verwenden und gleichzeitig die Parameteranzahl zu reduzieren. [Young & Woodland 93]

Beim State-Tying sollen die Zustände akustisch ähnlicher Triphone zusammengefasst werden, sodass sie Emissionswahrscheinlichkeiten teilen. Es gibt das datengetriebene Bottom-Up-Verfahren, und das Top-Down-Verfahren mit Entscheidungsbäumen. Diese werden in dieser Arbeit verglichen und der Nutzen des State-Tyings verglichen mit Spracherkennung ohne

Tying untersucht. State-Tying kann auf alle kontextabhängigen Modelle angewandt werden, da Triphone die höchste Genauigkeit haben und am häufigsten zum Einsatz kommen, werden die Verfahren nur an ihnen erklärt. Für Biphone sind die Verfahren entsprechend mit nur einem Kontext.

2. State-Tying

„Das Ziel des State-Tying ist es, die Anzahl der Parameter des Spracherkennungssystems zu reduzieren, ohne dabei signifikant die Genauigkeit des Modells herunterzustufen.“ (Beulen *et al.* 97, p. 1).
Wie in Tabelle 1 erkennbar, werden die Hälfte der Triphone im Trainingskorpus des Wall Street Journals, ein von ARPA zusammengestellter Spracherkennungskorpus, weniger als zehnmal gesehen, somit gibt es für sie kaum Trainingsdaten, wodurch sie später bei der Spracherkennung nicht gut erkannt werden. Außerdem kamen ca. 4% der Triphone gar nicht vor, weswegen ihnen kein akustisches Modell zugeordnet werden kann. Ohne State-Tying werden diese durch Monophone oder Backing-Off-Modelle modelliert, die eine Verallgemeinerung von Monophonen sind, dementsprechend wird die Kontextabhängigkeit nicht repräsentiert.

Häufigkeit im Trainingstext	Anzahl der Triphone
1-9	3737
10-19	1072
20-29	575
30-39	358
40-49	259
50-59	199
60-69	188
70-79	145
80-89	130
90-99	89
$\geq$ 100	1082
Summe	7834

Tabelle 1: Häufigkeiten von Triphonen im Trainingskorpus WSJ0 (K. Beulen 1999 p.38)

Unter Triphonen gibt es akustische Ähnlichkeiten, so hat im Deutschen beispielsweise ein ‚s‘ als rechtes Kontextphonem auf ein ‚e‘ weniger Einfluss als ein ‚i‘; w|e|n und k|e|s haben z.B. eine ähnliche Aussprache des Kernphonems ‚e‘ und können zusammengefasst werden.
Bei der Verknüpfung zweier Zustände werden die Emissionsverteilungen der einzelnen Zustände durch eine einzige Mischverteilung modelliert, wodurch für die Parameter der Mischverteilung mehr Trainingsdaten zur Verfügung stehen.

Beim State-Tying sollen die Triphonzustände, die akustisch ähnlich sind, zusammengefasst werden, dafür gibt es zwei verschiedene Verfahren: das Bottom-Up- und das Top-Down-Verfahren. Beide verwenden ein Abstandsmaß, welches ein Maß für die Ähnlichkeit zwischen Beobachtungen in einem Merkmalsraum ist. Dabei werden Beobachtungen, deren Abstand geringer als zu anderen ist, identifiziert und verknüpft. Diese durch ein akustisches Abstandsmaß verknüpften Beobachtungen bilden ein sogenanntes Cluster. [Beulen 99]
Das Abstandsmaß kann durch die Differenz der Log-Likelihood-Funktion von Gaußverteilungen bestimmt werden, welche die Wahrscheinlichkeit eines Triphons zu einem Sprachsignal angibt. Liegt eine niedrige Differenz der Log-Likelihood zweier Triphone vor,

sind sie zueinander akustisch ähnlich. Das Abstandsmaß kann anders bestimmt werden, beschrieben in [Kramer 96], was jedoch kaum Unterschiede bezüglich des Maßes und der Erkennungsrate ergibt. [Beulen 99]
Beide Verfahren sind Clusterverfahren, d.h. sie arbeiten mit einem Abstandsmaß und bilden Cluster, die am Ende verknüpfte Triphonzustände ergeben. Das Bottom-Up-Verfahren ist ein rein datengetriebenes Verfahren, bei dem durch einen agglomerativen Algorithmus mehrere kleine Cluster miteinander zu einem größeren verknüpft werden, während beim Top-Down-Verfahren mit einem großen Cluster begonnen wird, das in kleinere aufgespalten wird, indem es einen Entscheidungsbaum mit phonetischen Fragen durchwandert.

Zur Entstehung des State-Tyings:
1988 wurden von [Lee 88] generalisierte Triphone eingeführt, die im Gegensatz zum State-Tying Triphone als Ganzes miteinander verknüpfen und die Segmente sich jeweils eine Mischverteilung teilen, nicht gesehene Triphone wurden auf Backing-Off-Modelle abgebildet. Die Wortfehlerrate konnte mit dem ‚Ressource Management'-Task von 4,6% auf 4,2% gesenkt werden.
1992 wurde das State-Tying von [Hwang *et al.* 92] eingeführt, wobei für die Modellierung der ungesehenen Triphone ebenfalls Backing-off-Modelle verwendet wurden, da es vorerst nur ein Bottom-Up-Verfahren gab.
Das Verfahren wurde von [Young *et al.* 94] und [Hwang 93] unabhängig voneinander mit Entscheidungsbäumen erweitert.

2.1 Bottom-Up-Verfahren

Beim Bottom-Up-Verfahren werden Cluster durch Zusammenführen von Datenpunkten erstellt, die zuerst ihr eigenes Cluster darstellen und mit einem anderen Cluster zu einem neuen verschmolzen werden. Es wird solange wiederholt, bis die gewünschte Anzahl an Clustern erreicht ist. Das geschieht anhand des Abstandsmaßes, zwei Cluster werden so ausgewählt, dass die Log-Likelihood-Differenz zwischen ihnen und dem neuen Cluster möglichst gering ist, da das bedeutet, dass sich die vorherigen und das neue Cluster sehr ähnlich sind. [Beulen 99]

Das Bottom-Up-Verfahren, beginnt mit nicht verknüpften Triphonzuständen, die zunächst Werte für ihre Parameter benötigen, damit das Abstandsmaß bestimmt werden kann, anschließend werden sie in zwei Phasen miteinander verknüpft.
Zuerst werden Phoneme trainiert, deren Parameter als Anfangsinitialisierung der kontextabhängigen Modelle dienen, indem sie an die entsprechenden Stellen kopiert werden. Anschließend werden sie erneut mit Baum-Welch trainiert und die Auftrittshäufigkeit gezählt. Das Ergebnis sind Triphone mit den vorherigen Phonemen als Kernphonem. [Young & Woodland 93]
Unter ihnen werden in der ersten Phase paarweise Triphone, deren Abstandsmaß möglichst klein ist, bestimmt und zusammengefasst. Das wird solange fortgeführt, bis das Abstandsmaß einen gewissen Schwellwert überschreitet. Der Schwellwert bestimmt, wie verschieden die Cluster am Ende sein sollen und dadurch auch, wie viele Cluster es geben wird. Je höher der Schwellwert, desto weniger Cluster gibt es und dementsprechend mehr Daten für jedes Cluster, ist er jedoch zu hoch, sinkt die Kontextabhängigkeit. Ein Schwellwert von unendlich entspricht Monophonen, da alle Triphone desselben Kernphonems zusammengefasst werden und keine Aussprache berücksichtigt wird, während ein Schwellwert von 0 nicht verknüpfte Triphone bedeutet. [Young & Woodland 93]
In der zweiten Phase werden alle entstandenen Zustände, deren Auftrittshäufigkeit unter

einem bestimmten Wert (meist 100) lag, mit ihren direkten Nachbarn verknüpft. Dieser Schritt ist dafür da, dass für alle Zustände genügend Trainingsdaten für die Schätzung der Emissionswahrscheinlichkeiten vorhanden sind.
Für das Verfahren kann der k-Means-Algorithmus angewendet werden, ein Algorithmus, der anhand des Abstandsmaßes die ähnlichsten Cluster heraussucht und verschmilzt, anschließend für alle mit zu geringem Auftritt direkte Nachbarn sucht und weiter verschmilzt.
Dieser Algorithmus funktioniert nicht in Top-Down-Richtung, da dort ungleichmäßige Cluster entstehen. [Young & Woodland 93]

Alle verknüpften Triphonzustände sind eine Mindestanzahl aufgetreten, somit liegen für sie Trainingsdaten vor, Triphonen, die nicht gesehen wurden, kann allerdings kein akustisches Modell zugeordnet werden, diese werden einfach durch Monophone bzw. Backing-Off-Modelle modelliert, somit ohne Berücksichtigung des Kontextes. [Hwang 93]

2.2 Top-Down-Verfahren

Es wird mit einem großen Cluster, das alle Triphonzustände enthält, angefangen und so aufgespalten, dass die akustische Verschiedenheit möglichst groß ist, die durch das Abstandsmaß bzw. die Log-Likelihood-Funktion bestimmt wird.

Die Aufteilung wird solange fortgeführt, bis das Abstandsmaß aller Verteilungen kleiner als ein Schwellwert wird.

2.2.1 Entscheidungsbäume

Ein phonetischer Entscheidungsbaum ist ein binärer Baum, an dessen inneren Knoten phonetische Fragen und an den Blättern Indizes einer Mischverteilung stehen.
Das Ausgangscluster beginnt an der Wurzel, an jedem Knoten kann sich das Cluster spalten, indem alle Triphone, die die Frage mit „Ja“ beantworten, die linke Verzweigung wählen und alle anderen die rechte. Die resultierenden Cluster an den Blättern bilden die verknüpften Triphonzustände. [Beulen 99]
Zu jedem Kernphonem wird ein Entscheidungsbaum mit phonetischen Fragen erstellt, alle Triphone dieses Kernphonems bilden ein Anfangscluster, das den Baum durchwandert und bei jeder Frage gespalten wird.

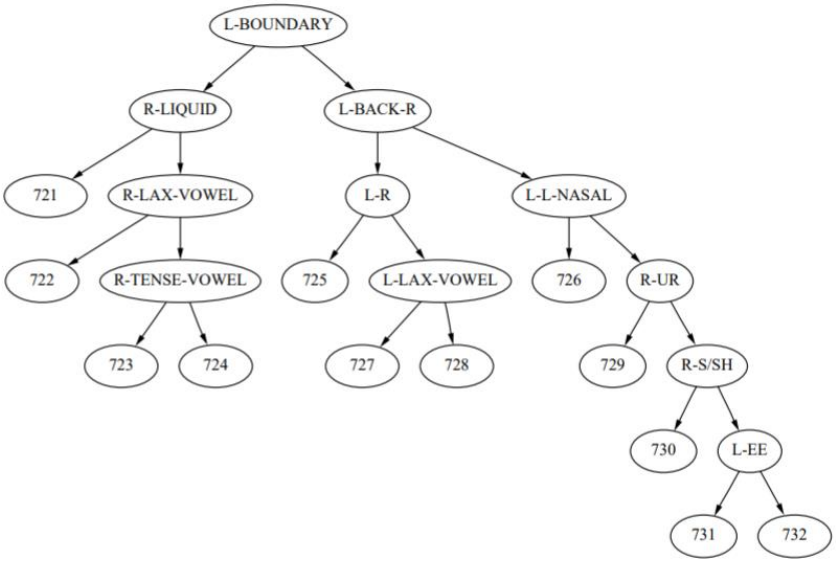

Abbildung 1: Entscheidungsbaum zum englischen Phonem „th“ (Beulen et al. 1997 p.2)

Am Beispiel des Entscheidungsbaumes in Abb. 1 werden die Triphone mit dem Kernphonem ‚th' in 12 Gruppen geteilt, denen jeweils eine Emissionsverteilung zugeordnet wird. Jedes Blatt sollte eine Mindestanzahl aufgetreten sein.

Phonetische Fragen haben die Form „Ist der linke/rechte Kontext aus der Menge X?", wobei X eine bestimmte Menge an Phonemen ist, also die phonetischen Klassen wie Vokale, Diphthong usw., z.B. „Ist der linke Kontext ein Diphthong?". [Young *et al.* 94]
Die Fragen müssen so gewählt werden, dass jede Verzweigung die Log-Likelihood maximal ansteigen lässt. Es wird fortgeführt, bis der Anstieg der Log-Likelihood eine gewisse Schwelle unterschreitet, an dem Punkt wird gestoppt.

Zum Generieren der Fragen für Entscheidungsbäume wird linguistisches Fachwissen benötigt, es gibt jedoch Rankings zu den nützlichsten Fragen für den Wall-Street-Journal, die den Anstieg der Log-Likelihood zu verschiedenen Phonemklassen in verschiedenen Knoten auflisten und das Bauen des Entscheidungsbaumes unterstützen. In der Liste zeigt sich, dass Vokale als rechtes Kontextphonem den größten Einfluss auf die Aussprache haben. [Young *et al*. 94]
[Beulen 99] beschreibt außerdem ein automatisches Verfahren zum Generieren der phonetischen Fragen für Entscheidungsbäume, bei dem die Fragen durch einen Algorithmus erzeugt werden.
Ein Entscheidungsbaum kann jedem beliebigen Triphonzustand ein akustisches Modell zuordnen, somit erhalten auch im Trainingskorpus ungesehene Triphone ein Modell und können mittrainiert werden.

Die Aufteilung der Menge durch den Entscheidungsbaum ist bezüglich der Log-Likelihood nicht immer optimal und an den verschiedenen Blättern können weiterhin akustisch ähnliche Zustände resultieren. In dem Fall kann der Entscheidungsbaum gekürzt werden, indem Knoten verschmolzen werden, die in einem geringen Anstieg der Log-Likelihood enden. Das kann entweder während des Bauens des Entscheidungsbaumes geschehen oder nachdem der Baum fertig gebaut ist, indem mit einem Bottom-Up-Verfahren Knoten mit dem geringsten Anstieg der Log-Likelihood entfernt und durch einen Elternknoten ersetzt werden. [Beulen 99]

3. Vergleich

3.1 Top-Down vs. Bottom-Up

System	Feb'89	Oct'89	Feb'91	Sep'92
Agg D-D	4.10	4.84	3.78	8.05
Tree	3.87	4.99	3.74	7.31

Tabelle 2: WER [%] Agglomerative data-driven vs. Tree-based, (Young et al. 1994 p.5)

Die Tabelle zeigt einen Vergleich der Wortfehlerrate mit derselben Triphonmenge und Trainingsmethode. Beide Spracherkenner verwendeten wortinterne Triphone und hatten ca. 1600 verknüpfte Zustände. Dieser Vergleich wurde mit dem Ressource-Management-Task von [Young *et al.* 94] durchgeführt.

Die Verwendung des „Agglomerative data-driven", also das Bottom-Up-Verfahren zeigt bei 3 von 4 Durchgängen eine geringfügig höhere Fehlerrate als die Verwendung eines phonetischen Entscheidungsbaumes. Beide Verfahren weisen eine ähnlich gute Performanz

auf, der größte Nachteil des datengetriebenen Verfahrens ist, dass ungesehenen Triphonen nur Backing-off-Modelle zugeordnet werden können, die meist nur eine ungenaue Annährung sind. [Beulen 99] Beim Top-Down-Verfahren kann diesen ein präzises akustisches Modell zugeordnet werden. [Young *et al.* 94]
Dass beide Verfahren trotz dessen eine ähnliche Performanz haben, ist dadurch zu erklären, dass Triphone, die in Trainingsdaten selten vorkommen meistens in der Sprache ebenfalls selten vorhanden sind, wenn die Trainingsdaten eine gute Repräsentation der Sprache sind. Deswegen haben sie keine allzu große Auswirkung auf die Erkennungsrate, allerdings sind Spracherkenner mit der zweiten Variante vor allem auf lange Sicht robuster. Außerdem wurde in Tabelle 2 mit wortinternen Triphonen trainiert, bei wortübergreifenden Triphonen steigt die Zahl der ungesehenen Triphone an.
Der Nachteil des Clusterns mit Entscheidungsbäumen ist, dass die Anzahl der phonetischen Fragen begrenzt ist und es weniger Aufteilungsmöglichkeiten gibt, in der Hinsicht ist das datengetriebene Verfahren genauer; tauchen die die ungesehenen Triphone nicht im Erkennungsvokabular auf, wenn also die Testdaten, anhand derer die WER berechnet wird, den Trainingsdaten entsprechen, weisen datengetriebene Verfahren in [Hon 92] eine bessere Erkennungsrate auf.

3.2 State-Tying vs. kein Tying

Tying	Zustände		Dichten	DEL-INS	WER
	vor	nach	(k)	[%]	[%]
kein Tying	2338	2338	246	–	8.8
datengetrieben	5698	2005	168	1.2-1.0	8.1
Entscheidungsbaum	23502	2001	192	1.2-0.7	7.6

Tabelle 3: WER [%] für State-Tying auf WSJ-5k (Beulen 1999 p. 58)

In dieser Tabelle unterscheiden sich die Fehlerraten mit Anwendung des datengetriebenen Verfahrens und des Entscheidungsbaumes um deutlich mehr als in der vorherigen Tabelle, das ist allerdings dadurch zu erklären, dass die Anzahl der Zustände beim Entscheidungsbaum deutlich mehr reduziert wurde. Während die Anzahl mit dem datengetriebenen Verfahren um 65% reduziert wurde, wurden sie beim Entscheidungsbaum um mehr als das elffache reduziert. Für einen Vergleich der beiden Verfahren sollte jedoch die gleiche Anzahl und Reduzierung betrachtet werden.

Durchs State-Tying zeigt sich eine relative Verbesserung der Wortfehlerrate von ca. 8,6%. Andere Vergleiche zeigten ähnliche Ergebnisse, so gab es bei [Young & Woodland 94] auf WSJ-Task und RM-Task mit wortübergreifenden Triphonen durchgeführt eine durchschnittliche relative Verbesserung von 14%. Außerdem zeigt sich in [Beulen 99], dass die WER sinkt, je größer der Faktor ist, um den die Anzahl der Triphonzustände reduziert werden, jedoch fehlt eine Angabe zum optimalen Wert, denn wie vorher beschrieben, würde eine zu hohe Reduzierung Monophonen entsprechen. Da es in jeder Sprache nur zwischen 10 und 80 Phoneme gibt, sind ca. 2000 verbleibende Zustände immer noch weit von einem Training nur mit Monophonen entfernt.

Alle Vergleiche zeigen eine Verbesserung der WER durch das State-Tying, hinzu kommt, dass durch Parameterreduktion die Berechnungskomplexität während der Erkennung verringert wird.

4. Fazit

Das Grundproblem der Verwendung von Triphonen in der Spracherkennung ist, dass Trainingsdaten begrenzt sind und nicht für jedes Triphon genügend Trainingsdaten vorliegen. Damit trotzdem ausschließlich Triphone verwendet werden können, mit denen die Erkennungsrate am besten ist, da sie die Koartikulation beider Nachbarphoneme berücksichtigen, kann mit der Methode des State-Tyings die Parameterzahl reduziert werden.

Die Zustände der akustisch ähnlichen Triphone werden geclustert und teilen sich dabei eine Emissionsverteilung, sodass diese zusammen trainiert werden können.
Das führt zum einen wie in verschiedenen Versuchen gezeigt, zu einer besseren Erkennungsrate und zum anderen nimmt die Berechnungskomplexität während der Spracherkennung ab.
Es wurden zwei verschiedene Clusterverfahren vorgestellt, die akustisch ähnliche Triphone anhand eines Abstandsmaßes clustern: das datengetriebene Bottom-Up-Verfahren und das Top-Down-Verfahren mit einem phonetischen Entscheidungsbaum.

Beide Verfahren führen zu einer signifikanten Verbesserung der Performanz und haben beide Vor- und Nachteile; so sind datengetriebene Verfahren beim reinen Clustering genauer, da sie nicht durch begrenzte Fragen eingeschränkt sind, ihr Nachteil ist jedoch, dass sie ungesehenen Triphonen kein Modell zuordnen können und diese durch ungenaue Backing-Off-Modelle modelliert werden.

Es hat sich außerdem gezeigt, dass ein größeres Zusammenfassen, das am Ende zu weniger Triphonen führt, zu einer besseren Erkennungsrate führt, dabei gibt es jedoch einen Grenzwert, ab dem sie wieder sinken würde, da weniger Koartikulationseffekte berücksichtigt werden. Wie weit zusammengefasst wird, wird durch Vorgabe des Abstandsmaßes bestimmt, der Schwellwert gibt an, wie akustisch ähnlich die Triphone maximal sein sollen, damit sie geclustert werden.

Literaturverzeichnis

[Lee 88] K.-F. Lee, „Large Vocabulary Speaker Independent Continuous Speech Recognition: The SPHINX System“ April 1988

[Hon 92] H.-W. Hon, „Vocabulary-Independent Speech Recognition: The VOCIND System“ 1992

[Hwang *et al.* 92] M.Y. Hwang, X. Huang, F. Alleva, “Predicting Unseen Triphones with Senones,” 1992

[S. J. Young & P. C. Woodland 93] S. J. Young, P. C. Woodland 93 „The Use of State-Tying in Continuous Speech Recognition“ 1993

[Hwang 93] M.Y. Hwang, “Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition,” 1993

[S. J. Young *et al.* 94] S. J. Young, J. J. Odell, P. C. Woodland „Tree-Based State Tying for High Accuracy Acoustic Modelling“ 1994

[Kramer 96] M. Kramer „Bottom-Up-Clustering für Phonemmodelle in der Spracherkennung“ März 1996.

[K. Beulen *et al.* 97] K. Beulen, E. Bransch, H. Ney 97 „State Tying for Context Dependent Phoneme Models“ 1997

[K. Beulen 99] K. Beulen „Phonetische Entscheidungsbäume für die automatische Spracherkennung mit großem Vokabular“ 1999

[W. Reichl & W. Chou 2000] Wolfgang Reichl, Wu Chou „Robust Decision Tree State Tying for Continuous Speech Recognition“ 2000