KONFIDENZBERECHNUNG FÜR AUTOMATISCHE LABELS

Frank Duckhorn, Rüdiger Hoffmann

Institut für Akustik und Sprachkommunikation, TU Dresden
frank.duckhorn@tu-dresden.de

Kurzfassung: Die automatische Generierung von zeitlichen Markierungen, sogenannten Labels, ist eine häufige Anwendung in der akustischen Mustererkennung. Sie wird zum Beispiel verwendet, um Sprachdatenbasen für das Training eines Spracherkenners oder für die Sprachsynthese zu annotieren. Aber auch bei technischen Signalen soll in vielen Fällen der Zeitpunkt bestimmter Ereignisse automatisch detektiert werden. Um dies zu erreichen, wird ein akustischer Mustererkenner verwendet und die zeitliche Referenz in dem akustischen Signal protokolliert, bei dem der Erkenner zwischen bestimmten Zuständen wechselt. Die Genauigkeit der Position der Labels hängt sehr stark von den verwendeten Daten sowie der Modellierbarkeit des gesuchten Ereignisses ab. Für Sprachsignale ist die Qualität meist ausreichend, einen guten Spracherkenner vorausgesetzt. Bei technischen Signalen ist die Bandbreite deutlich größer.

Dieser Beitrag stellt ein Verfahren vor, welches den automatischen Labels eine Konfidenz zuordnet. Diese ermöglicht, die Genauigkeit eines Labels einschätzen zu können. Anhand von Anwendungsfällen wird deutlich, wie nützlich die Konfidenzberechnung für automatische Labels sein kann, um fehlerhafte Ergebnisse zu vermeiden.

1 Einleitung

Ein wichtiges Anwendungsgebiet der akustischen Mustererkennung ist die automatische Generierung von zeitlichen Markierungen. Das Ziel ist, den Zeitpunkt von bestimmten akustischen Ereignissen zu bestimmen oder die Grenze zwischen zwei akustisch unterschiedlichen Bereichen zu finden. In der Sprachverarbeitung wird dieses Verfahren benutzt um große Sprachdatenbasen automatisch zu annotieren [1]. Die manuelle Annotation ist hier sehr zeitaufwändig. Annotierte Sprachdatenbasen werden für Sprachsynthese und -erkennung gleichermaßen benötigt. Bei der Verarbeitung von technischen akustischen Signalen soll in vielen Fällen der Zeitpunkt von bestimmten Ereignissen detektiert werden [9, 7]. Diese Markierungen werden häufig nach dem Englischen als Labels bezeichnet. Daher wird dieser Begriff auch hier verwendet werden.

Das Verfahren der automatischen Annotation ist jedoch nicht fehlerfrei. Die Genauigkeit der erzeugten Labels hängt davon ab, ob der verwendete Mustererkenner ausreichend trainiert ist sowie von der Unterscheidbarkeit der Ereignisse in den akustischen Signalen. Bei der Annotation von Sprachdatenbasen werden die automatisch erzeugten Labels deshalb häufig manuell korrigiert.

Dieser Beitrag stellt ein Verfahren vor, welches die automatisch erzeugten Labels bezüglich ihrer Genauigkeit oder Korrektheit klassifiziert. Dadurch können entweder für die manuelle Korrektur potentiell fehlerhafte Labels markiert werden oder fehlerhaft markierte Signale automatisch ausgesondert werden. Im Unterschied zu den meisten bisherigen Veröffentlichungen [2, 5, 3] wird die Position des Labels nicht optimiert, lediglich die Konfidenz der Labels wird

durch Mustererkennung automatisch bestimmt. Dazu werden Ergebnisse der automatisch detektierten Position ausgewertet und Merkmale bestimmt. Diese dienen im zweiten Schritt der Klassifikation der Konfidenz des Labels.

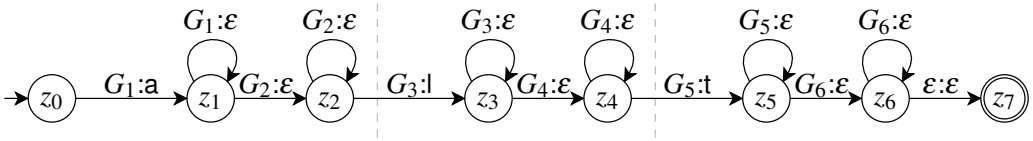
2 Erzeugung automatischer Labels mit endlichen gewichteten Automaten als Mustererkenner

Das hier vorgestellte Verfahren basiert auf der akustischen Mustererkennung mit gewichteten endlichen Automaten (Weighted Finite-State Transducer - wFST) [4, 8]. Um mit diesen Automaten automatische Labels erzeugen zu können, werden Modelle für die einzelnen Signalabschnitte, welche unterschieden werden sollen (in der Sprache zum Beispiel Phoneme), benötigt. Diese Modelle werden zum Beispiel in Form von Hidden-Markov-Modellen trainiert. Zu einem Signal, für welches die Folge von Modellen bekannt ist, können nun automatisch Labels erzeugt werden, indem die entsprechenden Modelle durch die Automatenoperation Produkt verkettet werden. Im entstandenen Automaten wird unter der gegeben Merkmalvektorfolge der Pfad mit dem kleinsten Gewicht bestimmt. Anhand der zeitlichen Position der Übergänge von einem Modell in ein anderes werden die Labels als Grenzen der Signalabschnitte definiert. Labels sind dabei nur auf dem Raster der Fenster der Merkmalextraktion möglich. Bei Signalen, für welche die Folge von Modellen nicht bekannt ist, kann der kleensche Abschluss der Summe aller Modelle verwendet werden. In beiden Fällen wird somit die Folge von Labels erzeugt, welche mit dem trainierten Modellen am besten übereinstimmt (im ersten Fall unter Berücksichtigung der bekannten Modellfolge). Abweichende Varianten werden nicht beachtet.

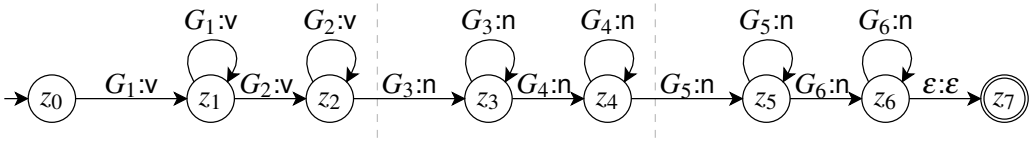
3 Bestimmung der Wahrscheinlichkeitsverteilung eines Labels

Die Bestimmung der Konfidenz eines Labels basiert auf der Wahrscheinlichkeitsverteilung dieses Labels. Diese Verteilung gibt an, wie wahrscheinlich sich das Label an einer bestimmten zeitliche Position befindet. Sie wird unter Berücksichtigung der Merkmalvektorfolge sowie des durch die trainierten Modelle erzeugten Automaten geschätzt.

Die Wahrscheinlichkeitsverteilung eines Labels wird ermittelt, indem zu jedem Label nicht nur der Pfad kleinsten Gewichtes durch den Automaten betrachtet wird, sondern zusätzlich eine bestimmte Anzahl von Pfaden nächst größeren Gewichtes. Pfade, welche das Label an die gleiche zeitliche Position setzten, werden dabei ignoriert. Um das zu erreichen wird der Automat für jedes Label modifiziert. Die Ausgabesymbole des Automaten werden geändert, so



(a) Phonemerkennungsautomat



(b) Modifizierter Automat

Abbildung 1 - Beispiel für Modifikation des Automaten zur Bestimmung der Wahrscheinlichkeitsverteilung des Übergangs vom Phonem *a* zu *l* in *alt*

dass der Automat zu jedem Zeitschritt genau ein Symbol ausgibt. Vor dem untersuchten Label wird immer das gleiche Symbol (v) ausgeben, danach ebenso, jedoch ein anderes Symbol (n). Somit unterscheidet sich die Ausgabesymbolfolge eines Pfades genau dann, wenn des Label an eine andere zeitliche Position gesetzt wird. In der Abbildung 1 ist ein Beispiel für diese Modifikation dargestellt. Im geänderten Automat wird nun nach den Pfaden mit dem geringsten Gewicht gesucht, welche sich aber in ihrer Ausgabesymbolfolge unterscheiden.

In der Tabelle 1 sind Beispiele für Gewichte von Pfaden für ein bestimmtes Label angegeben. Sie sind aufsteigen nach Gewicht $w(U_t)$ sortiert. Der Zeitpunkt t des Labels ist stets verschieden. Diese beiden Beispiele stammen aus der Kabelfehlerortung (siehe Abschnitt 5) und werden im Folgenden zur Erklärung des Verfahrens verwendet.

Zeitpunkt t [ms]	2.5	2.0	1.5	3.0	1.0	0.5	0.0	3.5	...
Pfadgewicht $w(U_t)$	3434	3437	3450	3455	3468	3484	3501	3507	...
Wahrscheinlichkeit $\tilde{p}(t)$	0.95	0.05	0.00	0.00	0.00	0.00	0.00	0.00	...
Wahrscheinlichkeit $\tilde{p}_{10}(t)$	0.48	0.35	0.09	0.06	0.02	0.00	0.00	0.00	...

(a) Beispiel 1

Zeitpunkt t [ms]	9.0	13.0	9.5	13.5	8.5	10.0	12.5	11.0	...
Pfadgewicht $w(U_t)$	2602	2604	2605	2607	2607	2608	2609	2609	...
Wahrscheinlichkeit $\tilde{p}(t)$	0.77	0.17	0.04	0.01	0.01	0.00	0.00	0.00	...
Wahrscheinlichkeit $\tilde{p}_{10}(t)$	0.08	0.07	0.06	0.05	0.05	0.05	0.04	0.04	...

(b) Beispiel 2

Tabelle 1 - Beispiele für Gewichte von Pfaden für ein bestimmtes Label sowie deren geschätzte Wahrscheinlichkeiten

Aus diesen verschieden Gewichten lässt sich eine Wahrscheinlichkeitsverteilung der Position des Labels schätzen. Die Verteilung ist diskret, da die Schätzung nur für die Position der Fenster der Merkmalextraktion möglich ist (im Beispiel mit Abstand 0.5ms). Durchgeführt wird die Schätzung durch Transformation der negativ logarithmischen Gewichte in eine Likelihood $LL(t)$, Gleichung (1), sowie anschließende Normierung auf die Summe der benutzen Likelihood's, Gleichung (2). Es ergibt sich die Wahrscheinlichkeitsverteilung $\tilde{p}(t)$. Die praktische Anwendung hat gezeigt, dass der entstehende Dynamikbereich der geschätzten Wahrscheinlichkeiten unnatürlich groß ist. Daher wird der Parameter γ zur Regulierung des Dynamikbereiches eingeführt, Gleichung (3). Somit ergibt sich nun endgültig die geschätzte Wahrscheinlichkeitsverteilung des Labels zu $\tilde{p}_\gamma(t)$. Für $\gamma = 1$ gilt: $\tilde{p}_1(t) = \tilde{p}(t)$.

$$LL(t) = \mathrm{e}^{-w(U_t)} \tag{1}$$

$$\tilde{p}(t) = \frac{\mathrm{e}^{-w(U_t)}}{\sum_{t_i} \mathrm{e}^{-w(U_{t_i})}} = \exp - \left(w(U_t) - \bigoplus_{t_i} \ln w(U_{t_i}) \right) \tag{2}$$

$$\tilde{p}_\gamma(t) = \frac{\mathrm{e}^{-w(U_t)/\gamma}}{\sum_{t_i} \mathrm{e}^{-w(U_{t_i})/\gamma}} = \exp - \left(\frac{w(U_t)}{\gamma} - \bigoplus_{t_i} \ln \frac{w(U_{t_i})}{\gamma} \right) \tag{3}$$

Für die beiden Beispiele in Tabelle 1 sind die entstehenden Wahrscheinlichkeitsverteilungen $\tilde{p}(t)$ und $\tilde{p}_{10}(t)$ für $\gamma = 10$ angegeben. Außerdem ist Letztere in der Abbildung 2 dargestellt. Man sieht leicht, dass sich die Verteilungen hinsichtlich der Sicherheit der Position des Labels deutlich unterscheiden. Diese Sicherheit wird im Folgenden in Form der Konfidenz automatisch auf Grundlage der Wahrscheinlichkeitsverteilungen klassifiziert.

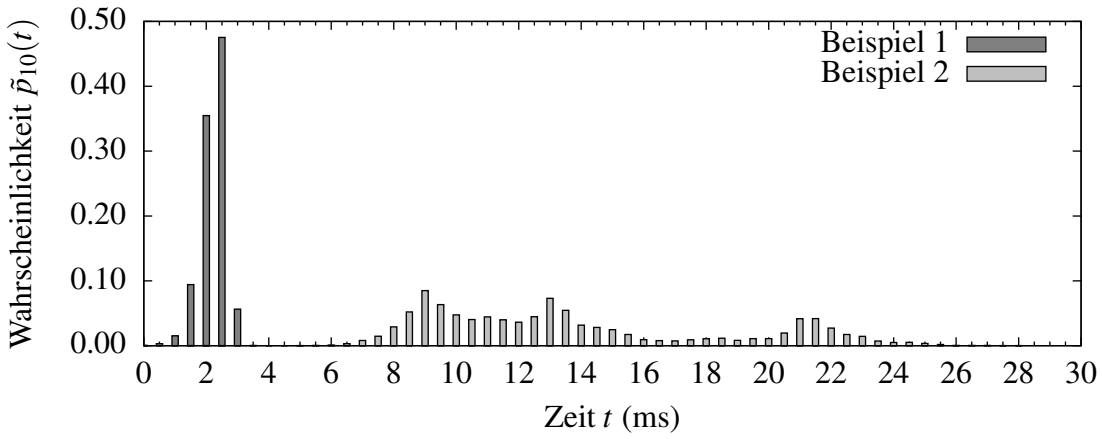


Abbildung 2 - Wahrscheinlichkeitsverteilungen der zeitlichen Position für die beiden Beispiele

4 Konfidenzklassifikation

Auf Basis der Wahrscheinlichkeitsverteilung eines Labels soll nun die Sicherheit oder Konfidenz dieses Labels klassifiziert werden. Dass heißt, es soll bestimmt werden, wie wahrscheinlich die automatisch detektierte Position des Labels korrekt ist. Dafür werden für die Wahrscheinlichkeitsverteilung verschiedene statistische Parameter bestimmt (siehe Tabelle 2), welche dann als Merkmale für einen Klassifikator in Form einer Support Vector Machine (SVM) [6] benutzt werden. Die SVM wird mit Daten trainiert, bei denen bekannt ist, ob die automatisch detektierte Position korrekt war oder nicht. Anschließend kann für unbekannte Daten eine Klassifikation durchgeführt werden.

	Merkmal		Berechnung	Werte für Beispiele	
1.	Varianz	=	$\mu_2^{(c)}(t)$	0.48	22.83
2.	Maximum	=	$\max(\tilde{p}_\gamma(t))$	0.17	0.08
3.	(Kurtosis)	=	$\mu_4(t)/\mu_2(t)^2$	1.12	1.50
4.	(Schiefe)	=	$\mu_3(t)/\mu_2(t)^{1.5}$	1.04	1.18
5.	Entropie	=	$-\sum_t \tilde{p}_\gamma(t)\log_2 \tilde{p}_\gamma(t)$	1.72	4.83

dezentraler k'ter-Moment : $\mu_k(t) = E(t^k) = \sum_t \tilde{p}_\gamma(t)\cdot t^k \;/\; \sum_t \tilde{p}_\gamma(t)$

zentraler k'ter-Moment : $\mu_k^{(c)}(t) = E((t-\mu_1(t))^k) = \sum_t \tilde{p}_\gamma(t)\cdot (t-\mu_1(t))^k \;/\; \sum_t \tilde{p}_\gamma(t)$

Tabelle 2 - Merkmale zur Konfidenz-Klassifikation mit den Werten für die beiden Beispiele aus den vorhergehenden Abschnitt

Die Merkmale wurden bezüglich ihrer Relevanz für die Klassifikation der Konfidenz ausgewählt. Eine Verteilung bei der die Sicherheit des Labels hoch ist, sollte folgende Eigenschaften haben:

- geringe Varianz (oder schmale Verteilung wie im Beispiel 1)
- großes Maximum (setzt sich gegenüber anderen Werten ab)
- hohe Kurtosis (steilgipflige Verteilung)

- geringe Schiefe (möglichst Symmetrie)
- geringe Entropie (Informationsgehalt als Quelle sollte klein sein)

Bei der Kurtosis wie auch bei der Schiefe haben Experimente gezeigt, dass die Klassifikationsleistung größer wird, wenn nicht die zentralen Momente verwendet werden. Deswegen folgt das dritte und vierte Merkmal nicht genau der Definition der Kurtosis beziehungsweise der Schiefe einer Verteilung. Es werden die dezentralen Momente zur Berechnung verwendet.

5 Anwendung in der Kabelfehlerortung

Gefördert durch:
Bundesministerium für Wirtschaft und Technologie
aufgrund eines Beschlusses des Deutschen Bundestages

Die Untersuchung dieses Anwendungsgebietes wurde durch die Zusammenarbeit mit der SebaKMT GmbH ermöglicht. Das Ziel der Kabelfehlerortung ist, die Positionsbestimmung von Kabeldefekten an erdverlegten Energieversorungsleitungen [7]. Es wird kurzzeitig eine hohe Spannung an das Kabel angelegt, so dass an der Fehlerstelle ein Überschlag auftritt. Die Entfernung zur Fehlerstelle wird dann Anhand der zeitlichen Differenz von magnetischem und akustischem Impuls bestimmt.
Die Aufgabe der Mustererkennung ist dabei die Bestimmung des Zeitpunktes des Impulses im akustischen Signal. Je größer die Entfernung zur Fehlerstelle ist, desto geringer ist der Pegel des Impulses und desto stärker wird er von Störgeräuschen überlagert. Um die Position des Impulses automatisch zu detektieren werden auf Basis von manuell annotierten Signalen Hidden-Markov-Modelle trainiert. Die Klassifikationsleistung ist jedoch aufgrund der starken Störsignale bei großen Entfernungen nicht zufriedenstellend. Die automatisch klassifizierte Konfidenz der Impulsposition ermöglicht fehlerhaft erkannte Impulspositionen zurück zu weisen und so die Klassifikationsleistung zu erhöhen.
In der untersuchten Anwendung wurden insgesamt 4877 Aufnahmen von akustischen Signalen von sechs verschiedenen Kabelfehlern verwendet. Die Aufnahmen wurden in 14 verschiedenen Entfernungen bis zu 15 Meter durchgeführt. Alle Aufnahmen werden so geschnitten, dass sie mit dem Eintreffen des magnetischen Impulses beginnen. Durch Kreuzvalidierung wurde unter Auslassung je eines Kabelfehlers ein Erkenner für die Impulsposition trainiert und mit den, nicht zum Training verwendeten, Daten evaluiert. In der Abbildung 3 sind die Histogramme der erkannten Impulsposition bei unterschiedlichen Abständen für einen Kabelfehler dargestellt. Je größer die Entfernung zum Kabelfehler ist, desto größer muss auch der Abstand zwischen magnetischem und akustischem Impuls werden. An der Gesamthöhe der Balken im Histogramm sieht man, dass mit steigender Entfernung die erkannten Impulspositionen größer werden, jedoch werden auch immer mehr Impulse an offensichtlich falschen Positionen erkannt.
Die Färbung der Histogrammbalken in der Abbildung 3 stellt das Ergebnis der Konfidenzklassifikation dar. Der verwendete SVM-Klassifikator wurde mit den zwei Klassen „zurückgewiesen“ und „akzeptiert“ trainiert. Ein Impuls wurde, je nach dem ob er im manuell annotierten Bereich erkannt wurde oder nicht, einer der beiden Klassen zugeordnet. Bei der Evaluation wurden die Impulse die nahe der Trenngerade der SVM lagen in die Klasse „unentschieden“ eingeteilt.
Die Tabelle 3 zeigt, dass die Erkennungsrate der Mustererkennung in der Kabelfehlerortung für die benutzen Daten nur bei 36,9% liegt. Indem nur die Impulse betrachtet werden, welche von der Konfidenzklassifikation akzeptiert werden, kann die Erkennungsrate auf 58,9% gesteigert werden. Die Erkennungsrate der Konfidenzklassifikation bei ausschließlicher Verwendung der Klassen „akzeptiert“ und „zurückgewiesen“ liegt bei 70,2%.

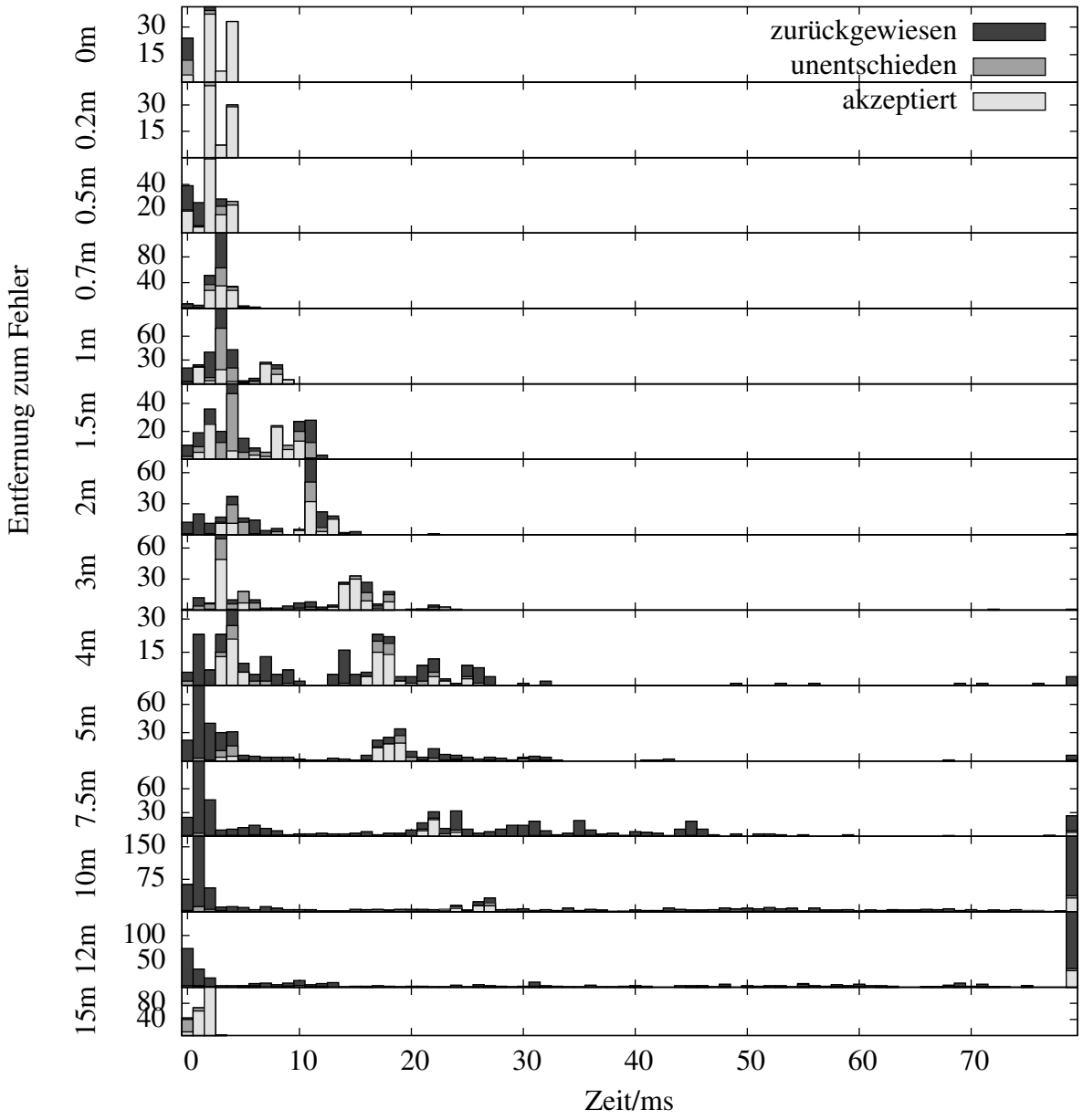


Abbildung 3 - Histogramm der erkannten Impulsposition für die Kabelfehlerortung bei verschiedenen Abständen zur Fehlerstelle. Die Färbung der Histogrammbalken stellt das Ergebnis der Konfindenzklassifikation dar.

	absoult		relativ	
	richtig	falsch	richtig	falsch
akzeptiert	791	551	**58,9%**	41,1%
unentschieden	439	266	62,3%	37,7%
zurückgewiesen	572	2258	20,2%	79,8%
Summe	1802	3075	**36,9%**	63,1%

Tabelle 3 - Ergebnisse der Konfidenzklassifikation in der Kabelfehlerortung

6 Anwendung in der automatischen Phonemmarkierung von Sprachsignalen

Eine zweites Anwendungsgebiet des hier vorgestellten Verfahrens findet sich in der Phonemmarkierung von Sprachsignalen. Dabei werden zu einem Sprachsignal mit bekannter Folge von Phonemen die Grenzen dieser Phoneme gesucht. Ziel der Konfidenzbestimmung ist eine Angabe, ob eine automatisch gefunden Grenze korrekt ist.

In der Abbildung 4 ist das Signal des Wortes „reliable“ dargestellt. Oberhalb des Signales sind die manuell markierten und unterhalb die automatisch markierten Grenzen der Phoneme zu sehen. Letztere können bedingt durch die Fenster bezogene Merkmalextraktion nur in einem festen Raster liegen (5ms). Neben den automatischen Grenzen ist deren Wahrscheinlichkeitsverteilung dargestellt, berechnet nach dem Verfahren im Abschnitt 3. Es wird deutlich, dass bei richtig erkannten Grenzen (I→l und b→l=) die Verteilung deutlich spitzgipfliger ist als bei falsch erkannten Grenzen (Anfang→r und aI→@). Daher kann die Wahrscheinlichkeitsverteilung zu einer großen Hilfe bei der manuellen Korrektur automatischer Grenzen werden.

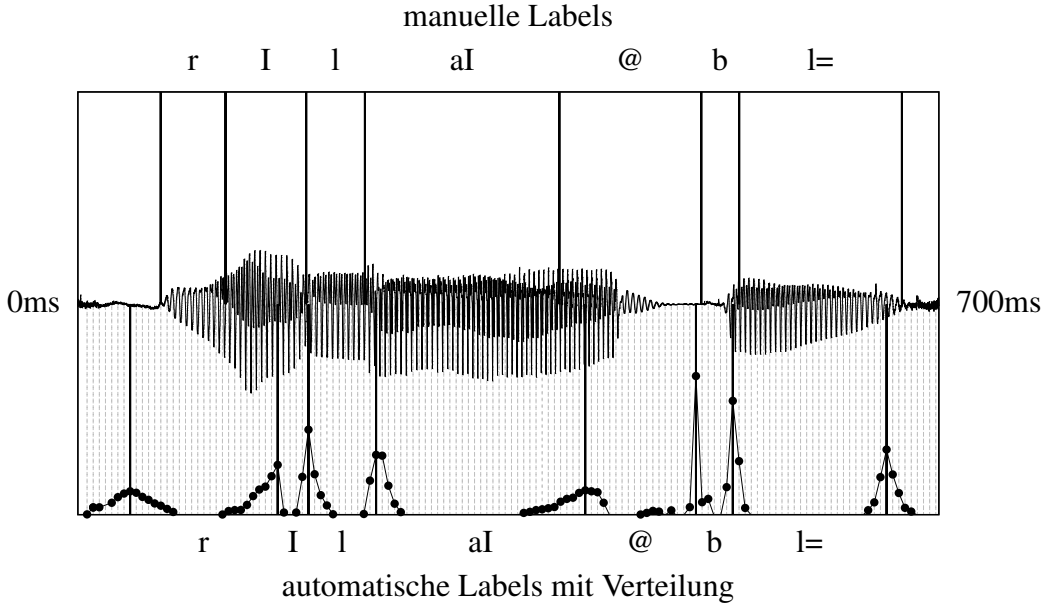


Abbildung 4 - Beispiel zur Wahrscheinlichkeitsverteilung von Phonemmarkierungen in Sprachsignalen (Wort: reliable)

Zur Klassifikation der Konfidenz werden zu jeder Phonemgrenze die in Tabelle 2 angegebenen Merkmale bestimmt. Außerdem werden fünf Werte vor und nach dem Maximum der Verteilung ebenfalls als Merkmale verwendet. Auf Basis dieser Merkmale wird mit einem Teil der Daten ein SVM Klassifikator trainiert, welcher zwischen richtig und falsch erkannten Phonemgrenzen unterscheiden soll. Grenzen werden als richtig eingestuft, falls der Abstand zur manuellen Referenz kleiner als 20ms ist.

Die Tabelle 4 zeigt die Ergebnisse der Konfidenzklassifikation. Die Evaluationsdaten enthalten 19469 Phonemgrenzen, von denen 82,5% richtig erkannt wurden (mit einem Abstand kleiner als 20ms). Davon wurden durch die Konfidenzklassifikation 72,7% akzeptiert und der Rest fehlerhaft markiert. Von den falsch erkannten Grenzen wurden 76,0% zurückgewiesen. Somit erhält man bei der automatischen Markierung von Phonemgrenzen in Sprachsignalen zusätzliche Information über potentiell falsch erkannte Grenzen. Die manuelle Korrektur der automatischen Grenzen wird erleichtert.

	Referenz	
	richtig	falsch
Gesamt	82,5%	17,5%
davon Konfidenz korrekt	72,7%	76,0%

Tabelle 4 - Ergebnis der Konfidenzklassifikation bei Sprachsignalen

7 Zusammenfassung

In diesem Beitrag wurde ein Verfahren vorgestellt, welches für automatisch erzeugte Markierungen Wahrscheinlichkeitsverteilungen der zeitlichen Position berechnen kann. Auf Basis dieser Verteilungen ist es möglich die Konfidenz der Markierung automatisch zu klassifizieren. Die Anwendung dieses Verfahrens in der Kabelfehlerortung zeigt, dass es möglich ist fehlerhafte Messungen zu detektieren und durch deren Ausschluss die Erkennungsleistung zu steigern. Die manuelle Korrektur von automatisch gesetzten Phonemgrenzen wird durch die Wahrscheinlichkeitsverteilung der Grenzen wie auch die Konfidenzklassifikation erleichtert.

Literatur

[1] KERI, V. und K. S. PRAHALLAD: *A comparative study of constrained and unconstrained approaches for segmentation of speech signal*. In: *Interspeech 2010*, Makuhari, Japan, Sep 2010.

[2] KOMINEK, J., C. BENNETT und A. BLACK: *Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis*. In: *Eurospeech 2003*, Geneva, Switzerland, 2003.

[3] LO, H.-Y. und H.-M. WANG: *Phone boundary refinement using ranking methods*. In: *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, S. 488 –492, 29 2010-dec. 3 2010.

[4] MOHRI, M., F. PEREIRA und M. RILEY: *Speech recognition with weighted finite-state transducers*. In: BENESTY, Y. J. und M. SONDHI (Hrsg.): *Handbook of Speech Processing*, S. 559–582. Springer, 2008.

[5] OGBUREKE, U. K. und J. CARSON-BERNDSEN: *Improving initial boundary estimation for HMM-based automatic phonetic segmentation*. In: *INTERSPEECH*, S. 884–887, 2009.

[6] OSUNA, E., R. FREUND und F. GIROSI: *Support Vector Machines: Training and Applications*. Techn. Ber., Cambridge, MA, USA, 1997.

[7] WITTENBERG, S. und F. DUCKHORN: *Abschlussbericht Projekt: Modulares Sensorsystem zur Positionsbestimmung von Kabeldefekten an erdverlegten Energieversorungsleitungen*. Techn. Ber., Institut für Akusitk und Sprachkommunikation, TU Dresden, 2011.

[8] WOLFF, M.: *Akustische Mustererkennung*. TUDpress, Dresden, 2011.

[9] WOLFF, M., U. KORDON, H. HUSSSEIN, M. EICHNER, R. HOFFMANN und C. TSCHOPE: *Auscultatory Blood Pressure Measurement using HMMs*. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Bd. 1, S. I–405 –I–408, april 2007.