

Segmentierung von TV-Werbespots für Automatische Spracherkennung und Jingle-Erkennung

Bernhard Rettenbacher, Franz Graf

*Joanneum Research Forschungsgesellschaft mbH, Institut für angewandte Systemtechnik, 8010 Graz, Österreich,
Email: bernhard.rettentbacher@joanneum.at*

Einleitung

TV Werbespots enthalten Sprache, Musik, Sound-Effekte und Hintergrundgeräusche. Besonders bei Werbespots treten diese Signalklassen oft simultan auf. Ist dies der Fall, verschlechtert sich beispielsweise die Erkennungsleistung eines automatischen Spracherkennungssystems. Die Erkennungsleistung kann verbessert werden, indem die akustischen Modelle entsprechend der akustischen Situation adaptiert werden. Einen wesentlichen Beitrag dazu liefert eine temporale akustische Segmentierung.

Frühere Arbeiten, die sich mit der Segmentierung von Sprache und Musik beschäftigten (z.B. E. Scheirer und M. Slaney [1]), widmeten sich dem Problem der Klassifizierung von bekannten homogenen Segmenten als Sprache oder Musik. Diese Arbeiten konzentrierten sich mehr auf die Extraktion von charakteristischen Merkmalen für die Klassifikation. Eine Einschränkung dieser Verfahren ist, dass die Systeme nur für Audiosegmente, die entweder Sprache oder Musik, nicht aber Klassen-Mixturen enthalten, evaluiert wurden.

Es gibt aktuellere Arbeiten, die Mixturen berücksichtigen. J. In [2] (Ajmera et al.) werden aus den A-Posteriori-Wahrscheinlichkeiten eines Spracherkenners die Merkmale „Entropy“ und „Dynamism“ extrahiert und eine Segmentmodellierung mit Hidden Markov Modellen und anschließender Klassifikation der Emissionswahrscheinlichkeiten durchführt. Der Ansatz von J. Pinquier et. al. [3] behandelt Sprache und Musik durch unabhängige Klassifikatoren, die durch eine Fusionsstufe kombiniert werden. Die Segmentierung wird für Sprach-, Musik- und auch Mixturen-Segmente durchgeführt

Im EU-Projekt „MediaCampaign“ [4] wird ein System zur automatischen medienübergreifenden Erkennung und Beobachtung von Werbe-Kampagnen in TV, Presse und Internet entwickelt. Der hier vorgestellte Prototyp dient der temporalen Segmentierung des Audiodatenstroms von TV-Werbespots. Ziel dieser Segmentierung ist, den Audiodatenstrom in homogene Segmente der Signalklassen „Sprache“, „Musik“, „Stille“ und „Sonstige Geräusche“ und in Segmente, die Mixturen aus den zuvor genannten Signalklassen enthalten, zu zerlegen, und diese den nachfolgenden Modulen zur Automatischen Spracherkennung und zur Erkennung von Jingles zur Verfügung zu stellen. Der hier präsentierte Ansatz basiert auf der Verwendung von unabhängigen Klassifikatoren für jede Signalklasse und anschließender Fusionierung.

Materialien und Methoden

Der Segmentierer besteht aus einer Merkmalsextraktion, Klassifikatoren für jede Signalklasse, einer anschließenden Fusionsstufe und einem Segment-Generator. Der Segmentierer ist datenstrombasiert und kann daher sowohl zur Verarbeitung von Audiodateien als auch Audiodatenströmen eingesetzt werden.

Für die Merkmalsextraktion werden aus dem Audiodatenstrom Analysefenster mit einer Länge von 25 ms extrahiert. Die Fenster werden mit 60% Überlappung generiert. Aus diesen Fenstern werden spektrale, cepstrale und temporale Merkmale extrahiert. Anschließend werden die Merkmalsvektoren der Analysefenster zu 60% überlappenden Texturfenstern mit einer Fensterlänge von 250 ms zusammengefasst und zentrale Momente der Merkmale und deren Ableitungen berechnet.

Die extrahierten Merkmale bilden Merkmalsvektoren, die den Klassifikatoren präsentiert werden, wobei diese für jeden Klassifikator unterschiedlich zusammengesetzt sind. Für die Klassen „Sprache“, „Musik“ und „Sonstige Geräusche“ werden Gaussian-Mixture-Model-Klassifikatoren eingesetzt, die als 2-Klassen-Modelle mit Signalen der jeweiligen Klasse (positiv) und den restlichen Signalen des Trainingskorpus (negativ) trainiert werden. Signale der positiven Klasse enthalten neben den Signalen, die ausschließlich der Klasse zugeordnet werden können, Signale die Mixturen mehrerer Klassen enthalten.

Die Detektierung von Stille basiert auf der Berechnung der psychoakustischen Lautheit. Unterschreitet die Lautheit einen dynamisch berechneten Schwellwert, so wird das Fenster als Stille klassifiziert.

Die Ergebnisse der Klassifikatoren werden in einer Fusionsstufe geglättet und kombiniert. Die Glättung erfolgt durch gleitende Mittelwertbildung der A-posteriori-Wahrscheinlichkeiten der Gaussian Mixture Models, um einzelne Ausreißer zu entfernen. Eine Entfernung von einzelnen Ausreißern ist sinnvoll, da durch die 60%-Überlappung der Fenster die benachbarten Fenster das Ausreißer-Fenster vollständig überlappen. Danach werden die Ergebnisse der einzelnen Klassifikatoren kombiniert und anschließend durch die Anwendung der oben genannten Ausreißerbehandlung auf die diskreten Ergebnisse, ein weiteres Mal geglättet. Danach werden aus den fusionierten Klassifikationsergebnissen Segmente gebildet. Diese Segmente werden im MPEG-7 Datenformat ausgegeben.

Für das Training und die Evaluierung des Segmentierers wurden Werbespots in deutscher, holländischer und englischer Sprache aufgezeichnet. Die Signalqualität ist typisch

für analoge TV-Audiosignale mit auftretenden Frequenzen bis etwa 15-16 kHz. Bis jetzt wurden 280 Werbespots manuell annotiert. Die Annotation wurde in MPEG-7 und mit Hilfe eines Mehrspur-Audioeditors durchgeführt. Diese Annotation erfolgte detaillierter als es die Segmentierungsaufgabe erfordert, da so eine Zusammenstellung der Trainings- und Testkorpora auf Metadaten-Ebene möglich ist. So können etwa Korpora für Klassen mit und ohne Mixturen erstellt und miteinander verglichen werden. Auch eine nachträgliche Änderung der Klassendefinitionen ist so leicht möglich. Die Zusammenstellung des Trainingskorpus erfolgt durch Abbildung der detaillierten Segment-Annotationen auf die Klassen des Segmentierers.

Resultate

Die Evaluierung des Segmentierers erfolgte für die einzelnen Klassifikatoren und die Fusionsstufe durch Kreuzvalidierung mit 10 Teilmengen. Als Messgrößen für die Klassifikatoren und die Fusionsstufe wird die Korrektklassifikationsrate (*accuracy*), Sensitivität (*recall*) und Relevanz (*precision*) angegeben.

Der verwendete Signalkorpus enthält 280 manuell annotierte TV-Werbespots mit einer Dauer von 20-60 Sekunden pro Spot. Die Signale wurden mit 44,1 kHz und 16 Bit abgetastet. Für Experimente zur Robustheit des Klassifikators wurde der Korpus zuerst mit 12 kHz, 8 Bit requantisiert, dann wieder in das Ursprungsformat zurückgewandelt.

Tabelle 1 zeigt die Ergebnisse für die einzelnen Klassifikatoren, Tabelle 2 die Ergebnisse nach der Fusion. Die Korrektklassifikationsrate für die Fusion und daher für den gesamten Segmentierer beträgt 74,7 %. Der Klassifikator für „Sonstige Geräusche“ befindet sich noch in einem frühen Stadium, weshalb ausschließlich Ergebnisse der Klassen „Sprache“, „Musik“ und „Stille“ und Mixturen von „Sprache“ und „Musik“ angegeben werden.

Tabelle 1: Ergebnisse der einzelnen Klassifikatoren

	Korrekt-klassifikationsrate	Sensitivität	Relevanz
Sprache	85,1 %	87,2 %	90,2 %
Musik	84,8 %	95,6 %	86,9 %

Tabelle 2: Ergebnisse der Klassifikator-Fusion

	Sensitivität	Relevanz
Sprache	33,1 %	60,5 %
Musik	78,0 %	71,9 %
Stille	73,9 %	63,1 %
Sprache + Musik	83,4 %	78,3 %

Die Evaluierung des requantisierten Korpus zeigte, dass nach erneutem Training, identische Ergebnisse erzielt werden können.

Diskussion

Der Ansatz, unabhängige Klassifikatoren für jede Klasse mit anschließender Fusionsstufe zu verwenden und anschließend

eine Kombination der optimierten Klassifikationsergebnisse vorzunehmen, ist einem Ansatz, bei dem alle Klassen und die Mixturen der Klassen durch einem einzigen Klassifikator klassifiziert werden, überlegen. Gerade in Segmenten, die Mixturen mehrerer Klassen enthalten, kommt es bei solchen Klassifikatoren - je nach Dominanz einer Klasse in einem Textur-Fenster - zu häufigen Klassenwechseln. 2-Klassen-Klassifikatoren können durch Auswahl der Merkmale, der Fensterlängen und des Klassifikatormodells optimiert werden, wodurch sich bereits vor der Kombination der Klassifikator-Ergebnisse Segmente mit längerer Dauer ergeben, die leichter modelliert werden können.

Die Klassifikationsraten des Systems weisen noch nicht die von anderen Autoren präsentierten Werte auf, jedoch zeigte sich, dass dies unter anderem auf den zu kleinen Trainingskorpus zurückzuführen ist. Es zeigt sich bei der Fusion, dass Sprach-Segmente häufig der Klasse „Sprache + Musik“ zugeordnet werden. Als Ursache konnte eine Häufung von Fehlklassifikationen des „Musik“/„Nicht-Musik“-Klassifikators bei Sprachsegmenten festgestellt werden, die durch das Training nicht abgedeckt sind.

Zusammenfassung

Das hier vorgestellte System zur temporalen Segmentierung des Audiodatenstroms von TV-Werbespots erreicht selbst in dieser durch viele Klassen-Überlappungen geprägten Domäne eine gute Erkennungsleistung. Durch Erweiterung des Trainingskorpus, Optimierung der Merkmale und deren Auswahl sowie durch Optimierung der Klassifikatoren kann die Erkennungsleistung weiter erhöht werden. Das System soll weiters durch die Weiterentwicklung des Klassifikators für „Sonstige Geräusche“ vervollständigt werden. Zur Optimierung der Rechenperformance kann durch eine Verringerung der Abtastfrequenz eine Performanceoptimierung erreicht werden.

Dank

Die für das MediaCampaign Projekt durchgeführten Forschungs- und Entwicklungsarbeiten werden zu einem Teil von der Europäischen Union im 6. Rahmenprogramm mit dem Strategischen Ziel "Semantic-based knowledge and content systems" des IST Arbeitsprogrammes 2004 (IST FP6-027413) gefördert.

Literatur

- [1] Scheirer, E.; Slaney M.: Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. ICASSP, 1997.
- [2] Ajmera, J; McCowan, I; Bourlard, H: Speech/music segmentation using entropy and dynamism features in a HMM classification framework. Speech Communication. Vol. 40, no. 3, pp. 351-363. May 2003
- [3] Pinquier, J.; Senac, C.; André-Obrecht, R.: Speech and music classification in audio documents, ICASSP, 2002.
- [4] MediaCampaign homepage, URL: <http://www.media-campaign.eu>