1E2-03

音素片のカーネル主成分分析を用いたトピックセグメンテーション

Topic Segmentation Using Kernel Pricipal Component Analysis for Sub-Phonetic Segments

佐土原 健　　李 時旭　　児島 宏明
Ken Sadohara　Shi-wook Lee　Hiroaki Kojima

産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

This paper describes an open-vocabulary method for segmenting spoken documents into topically homogeneous blocks. Without transcribing a spoken document into text, the method builds topical clusters directly from a recognized sequence of Sub-Phonetic Segments (SPS), and thus it is not constrained in term of vocabulary or grammar. Each analysis interval constituting the clusters is considered as a high dimensional vector whose element is a frequency of any sequence of SPS. Then the kernel principal component analysis reduces the dimensionality by extracting principal components expected to be more related to topics. Using cosine similarity between analysis intervals represented by the principal components, the hierarchical clustering method forms topically homogeneous clusters. The presented method is shown to be effective by an experiment on topic segmentation of broadcast news.

1. はじめに

大量のマルチメディアコンテンツを容易に蓄積可能になった今日, 適切な索引化や構造化により, コンテンツの概要を素早く把握するための技術が切実に求められている. そして, そのための基礎的な技術として, コンテンツを意味的に等質な部分に分割するトピックセグメンテーションに関する研究が行われてきた. 本稿では, 音声に基づくトピックセグメンテーションに関して, 従来研究とは異なり, 音声認識によるキーワード抽出を用いない手法について報告する.

本手法では, 入力音声を, 通常の音素よりも粒度が細かく, 言語依存性の低い音素片 [3] の列として認識し, この音素片の列から直接トピックセグメンテーションを行う. 音素片は, 通常の音素よりも粒度が細かいので, 誤認識によって必要な情報が抜け落ちてしまう危険性が少ない. また, 一定長以下の任意の音素片列を分析の対象とすることにより, 固有名詞や, 省略語等の辞書に登録されていない未知語に基づくトピックセグメンテーションが可能になる.

まず, 音素片の列は, 一定幅の分析区間に分割され, 各分析区間に含まれる一定長以下の部分符号列の頻度を成分とするベクトルとして, 各分析区間を表現する. そして, このベクトルの余弦を類似性とする, 階層的クラスタリング法を用いて, 類似した分析区間をボトムアップに一つのクラスタにまとめることで, 音声のセグメンテーションを行う.

しかし, このようなベクトルは非常に高次元 [*1]であるので, これを直接取り扱うことは計算量的に困難である. また, ベクトルの余弦は意味的な類似性を反映しておらず, トピックに無関係な音素片列の影響を除去したり, トピックに共起する音素片列を一つの成分にまとめたりすることにより, より意味的な類似性を反映した低次元のベクトルに変換する.

このような, 次元縮小や基底の変換を目的として, 主成分分析がよく用いられるが, 変量の数が非常に多いので, 主成分分析をそのまま適用することは困難である. そこで, 本研究では, カーネル主成分分析 [1] を用いる. M 個の s 変量ベクトルの主成分分析は, s 行 s 列の共分散行列の対角化を行う必要があるが, カーネル主成分分析を用いると, 各ベクトルの内積を計算した M 行 M 列の $\boldsymbol{K}$ ($\boldsymbol{K}_{ij} = \langle \boldsymbol{x}_i \cdot \boldsymbol{x}_j \rangle$) の対角化により, 高々 M 個の主成分を効率良く計算することができる. また, 内積の計算も, 文字列カーネル [2] を用いると, 分析区間の長さに対して線形な計算量で計算可能である.

本稿では, このような語彙制約のないトピックセグメンテーション手法を提案すると同時に, ニュース音声を対象として, セグメント境界の精度に関する評価実験の結果を報告する. 実験結果は, 厳密なセグメント境界を得ることは難しいが, どの部分で, どのような内容が話されているかを把握可能な程度の, トピックセグメンテーションの可能性を示唆している.

2. 提案手法

音素片 [3] は, 調音結合を考慮した, 音素よりも粒度が細かい音声符号系の一種である. 例えば, 「神戸」が, XSAMPA 符号系において "`k o o b e`" と表記されるのに対して, 音素片によれば, "`#kcl kk ko ooo ob bcl bb be ee e#`" のように表記される. 文献 [4] は, 音声文献の検索に音素片を適用し, 音声を音素片に変換した上で, 検索音声に類似したデータベース中の音声を, 音素片列の距離計算に基づいて検索する手法を提案している. 言語依存性の低い音素片を用いることで, 辞書にない単語の検索が可能になるだけでなく, 粒度の細かさにより誤認識に対して頑健な検索が可能であることが示されている.

本論文で提案するトピックセグメンテーションにおいても, 入力音声は, まず音素片の列に変換される. 音素片列は, 図 1 のように, 一定幅の単位に分割され, 連続する ℓ 個の単位を一つの分析区間とする. そして, M 個の分析区間それぞれに対して, 長さ p 以下の音素片列の頻度を成分とするベクトル $\boldsymbol{x}_i$ ($1 \leq i \leq M$) を考える.

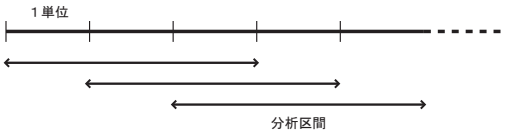

図 1: 分析区間

このベクトルの次元の縮小と基底の変換を行うために, 主成分分析を適用するが, ベクトルの次元が大きいため, 主成分分析を直接行うことは計算量的に困難であるので, カーネル主

連絡先: 佐土原 健, E-mail:ken.sadohara@aist.go.jp

*1 音素片の数を $|\Sigma|$, 分析対象とする音素片列の長さの最大値を p とすると, $O(|\Sigma|^p)$ 次元. 実験では, $|\Sigma| = 411$, $p = 20$.

成分分析 [1] を用いる. カーネル主成分分析を用いると, 内積を要素とするカーネル行列 $\boldsymbol{K}$ ($\boldsymbol{K}_{ij} = \langle \boldsymbol{x}_i \cdot \boldsymbol{x}_j \rangle$) を計算し, $\tilde{\boldsymbol{K}} = \boldsymbol{K} - \boldsymbol{U}\boldsymbol{K} - \boldsymbol{K}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{K}\boldsymbol{U}$ (ただし $\boldsymbol{U}_{ij} = \frac{1}{M}$) の対角化を行うことで, 主成分分析を行うことができる. $\boldsymbol{K}$ の計算には, 文字列カーネル [2, Remark 11.49] を用いることで, 分析区間 S_i, S_j の符号数をそれぞれ $|S_i|$, $|S_j|$ とするとき, $\boldsymbol{K}_{ij}$ を $O(p(|S_i| + |S_j|))$ で計算可能である.

このようにして得られた q 個の固有値 $\lambda^1 \geq \cdots \geq \lambda^q > 0$ と, これに対応する固有ベクトル $\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^q$ (ただし $\lambda^k \left\langle \boldsymbol{\alpha}^k \cdot \boldsymbol{\alpha}^k \right\rangle = 1$) を用いると, ベクトル $\boldsymbol{x}$ の 第-k ($1 \leq k \leq q$) 主成分への射影は, $\sum_{i=1}^{M} \alpha_i^k \langle \boldsymbol{x}_i \cdot \boldsymbol{x} \rangle$ で計算でき, 各分析区間は q 個の主成分を用いて, $[\boldsymbol{x}'_1 \cdots \boldsymbol{x}'_M]^\top = \boldsymbol{K} \left[\boldsymbol{\alpha}^1 \cdots \boldsymbol{\alpha}^q\right]$ と書ける. 結局, 高次元ベクトル $\boldsymbol{x}_i$ を陽に取り扱うことなく, その内積のみを文字列カーネルを用いて計算し, カーネル主成分分析を行うことで, より意味的な類似性を反映した低次元の分析区間ベクトル $\boldsymbol{x}'_i$ を得ることができる.

このようにして得られた分析区間ベクトルは, 隣接するベクトルどうしで $\ell - 1$ 単位分重複している. 次節で述べる実験では $\ell = 3$ であるので, 2 単位分の重複を以下のような補正で取り除いて, 各単位を $\boldsymbol{s}_i (1 \leq i \leq M+2)$ で表現する.

$$\boldsymbol{s}_i = \begin{cases} \boldsymbol{x}'_1 & i = 1 \text{ のとき} \\ \frac{\boldsymbol{x}'_1 + \boldsymbol{x}'_2}{2} & i = 2 \text{ のとき} \\ \frac{\boldsymbol{x}'_{i-2} + 2\boldsymbol{x}'_{i-1} + \boldsymbol{x}'_i}{4} & 2 < i < M+1 \text{ のとき} \\ \frac{\boldsymbol{x}_{M-1} + \boldsymbol{x}_M}{2} & i = M+1 \text{ のとき} \\ \boldsymbol{x}_M & i = M+2 \text{ のとき} \end{cases}$$

次に, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{M+2}$ に対して階層的クラスタリングを適用し, 与えられたクラスタ間距離を閾値として複数のクラスタに分割する. なお, 一つのクラスタの中に非連続な単位 s_i を含むこともあり得るが, その場合は, 連続する単位毎にクラスタを再分割する. また, クラスタ間の距離として, 以下のような標準的なクラスタ間距離 d を用いた.

$$d_0(\boldsymbol{s}_i, \boldsymbol{s}_j) = 1 - \frac{\langle \boldsymbol{s}_i \cdot \boldsymbol{s}_j \rangle}{\sqrt{\langle \boldsymbol{s}_i \cdot \boldsymbol{s}_i \rangle} \sqrt{\langle \boldsymbol{s}_j \cdot \boldsymbol{s}_j \rangle}}$$

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{m=1}^{|C_i|} \sum_{n=1}^{|C_j|} d_0(\boldsymbol{s}^i_m, \boldsymbol{s}^j_n)$$

ただし, $C_i = \{\boldsymbol{s}^i_1, \ldots, \boldsymbol{s}^i_{|C_i|}\} \subseteq \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{M+2}\}$.

3. 実験

本研究では, NHK の 15 分のニュース番組 6 回分を対象にして実験を行った. 各々の番組は, 人手により, 基本的に 1 文ごとに 1 単位に分割したが, 音素片の認識エンジンの実装上の都合により, 20 秒を超える文章は, 適当な無音区間で 20 秒未満に分割した. なお, 1 つの単位を 1 つの文章にしなければいけない本質的な理由は存在しないが, 今回の実験では, 以下で述べるようにトピックの境界を, どの程度正確に予測できるかを調べることを目的としているので, 1 単位の途中に境界が現われないように, このような分割を行った. 平均すると, 各番組の長さは 73.4 単位であり, トピックの数は 11.3 個, トピックの長さは 6.6 単位であった. なお, トピックの境界の判定については, ニュース映像のキャプションが変化した場所を, トピックの境界と考えた.

これらの番組に対して, 前節までに述べた手法を適用し, 得られるセグメント境界の再現率と適合率を評価した. 図 2 は,

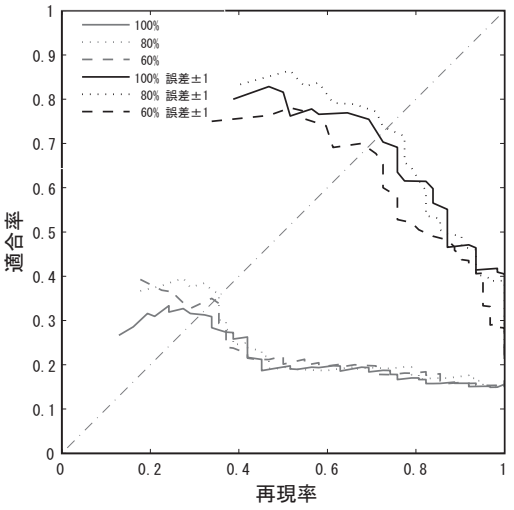

図 2: トピック境界精度

許容するクラスタ間の距離を変化させたときの, 再現率と適合率の変化の様子を示している. 灰色のカーブは, 境界を厳密に評価したときの精度で, 黒色のカーブは, トピック境界が, 前後に 1 単位ずれている場合でも正解としたときのカーブを示している. いずれの場合も, 累積寄与率が 100%, 80%, 60% になるように固有値の数を 3 通りに変化させてプロットしている.

4. おわりに

本稿では, 大語彙連続音声認識システム等によるキーワード抽出を行うことなく, 音声を, 音素よりも粒度の細かい音素片の列として認識した上で, 直接トピックセグメンテーションを行う手法を提案した. これにより, 一定長以下の任意の音素片列に基づいた, 語彙と文法に制約されないトピックセグメンテーションが可能になる. また, 本稿では, このような手法を用いて, ニュース音声のトピックセグメンテーションの実験を行い, トピック境界の精度を調べた. その結果, 厳密な境界を得ることは難しいものの, ある程度の誤差を許容して, どの部分で, どのような内容が話されているかをおおまかに把握可能な程度の, トピックセグメンテーションの可能性を示すことができた. 今後は, 教師ありトピックセグメンテーションとの融合を図りながら, さらなる精度向上を目指すと同時に, トピックの階層構造の正しさに関する評価も行いたい.

参考文献

[1] B. Schölkoph, et al.: "Nonlinear component analysis as a kernel eigenvalue problem", Neural Computation, **10**, 5, pp. 1299–1319 (1998).

[2] J.Shawe-Taylor and N.Cristianini: "Kernel methods for pattern analysis", Cambridge University Press (2004).

[3] K.Tanaka, et al.: "Speech data retrieval system constructed on a universal phonetic code domain", Proc. of ASRU, pp. 1–4 (2004).

[4] S. Lee, et al.: "Combining multiple subword representations for open-vocabulary spoken document retrieval", Proc. of ICASSP, pp.505–508 (2005).