FACULTAD DE INGENIERIA

Universidad de Buenos Aires

Departamento de electrónica

Plan de trabajo para Tesis de Grado

APRENDIZAJE BAYESIANO APLICADO A PROCESAMIENTO DEL HABLA DE CERO RECURSOS

Tesista: Battocchio Diego Javier
Padrón: 96695
Mail: diego.battocchio@gmail.com

Directora: Dra. Ing. Patricia Pelle
Mail: ppelle@fi.uba.ar

Índice

1. Objeto y área de la tesis

La presente tesis se enmarca en el área de procesamiento de señales, más específicamente en el procesamiento del habla y el objetivo de la misma es la implementación y puesta a prueba de un reconocedor de habla basado en aprendizaje bayesiano no supervisado que logre determinar, de manera automática, la estructura de un lenguaje, es decir, sus unidades acústicas elementales con las cuales se entrenará. Las técnicas utilizadas pertenecen a lo que es conocido en el ámbito del procesamiento del habla como aprendizaje de cero recursos, ya que no se basa en datos etiquetados, sino que utiliza solo las señales de habla, sin ningún tipo de información adicional sobre ellas. Este área se inspira en la idea de que los humanos aprenden su lenguaje sin la necesidad de otra cosa más que la constante exposición al mismo, por lo que formas de aprendizaje automático que solo cuenten con las señales de habla podrían ser técnicamente alcanzables.

2. Introducción

El reconocimiento automático del habla es un conjunto de tecnologías que permiten a una computadora la identificación de palabras pronunciadas por una persona a través de un micrófono. Para realizar esta tarea, se modelizan las palabras como cadenas ocultas de Markov conectadas de manera estadística según el tipo de lenguaje o de contexto lingüístico utilizado [3] y posteriormente se utiliza el algoritmo de Viterbi [1] para determinar la frase pronunciada. Cuando se busca que el reconocedor funcione para lenguajes completos, dada la cantidad infinita de palabras que existen en un vocabulario, la modelización de ellas se realiza con un paso extra desglosando las mismas en fonemas, que son finitos y forman todas las palabras de un vocabulario. De esta forma, es posible entrenar al algoritmo para reconocer cadenas de fonemas en vez de palabras. Para esto debe proveerse al reconocedor, en su etapa de entrenamiento, con las señales de habla, su transcripción a palabras y la composición fonética de cada una de esas palabras. Cuando se tiene toda esta información se dice que se esta entrenando de forma supervisada.

entrenar fonemas en lugar de palabras, y luego componer éstas con la correcta concatenación de los fonemas que las componen en la etapa de reconocimiento.

Esta forma de entrenamiento esconde dos hipótesis, la primera es que se tiene un diccionario fonético de las palabras a entrenar, y la segunda es que se posee la transcripción a palabras de los audios a utilizar. En una época en la que se tiene acceso a una cantidad masiva de audios sin etiquetar, el no poder usar estas señales por el hecho de que no se tenga su transcripción resulta una limitación importante. Es por eso que el objetivo de esta tesis es atacar los casos en los que estas hipótesis no se cumplen, entrenando al reconocedor de forma no supervisada.

3. Desarrollo previsto de la tesis

3.1. Teoría, enfoques y métodos a utilizar

3.1.1. Enfoque clásico

Como se adelantó en la sección precedente, los fonemas son modelizados como una cadena oculta de markov (HMM) con una cantidad de estados fija K, de topología izquierda-a-derecha y emisiones de mezclas de L gaussianas (GMM) [1, 3]. Esta estructura esta gobernada por un conjunto de parámetros Θ como son las medias, varianzas y pesos de las mezclas de gaussianas de las emisiones, la matriz de transición entre estados y las probabilidades iniciales. El entrenamiento clásico consiste en alimentar al reconocedor con señales de habla trascriptas a palabras en donde se conoce la composición fonética de cada una de ellas para que el reconocedor pueda entrenarse de forma supervisada, es decir, sabiendo que fonemas entrenar con cada señal. Esto se lleva acabo haciendo uso del algoritmo de Baum-Welch [2], obteniéndose así el conjunto de parámetros Θ que mejor describen las observaciones, y que serán usados posteriormente para el reconocedor.

3.1.2. Enfoque no supervisado

El enfoque clásico cuenta con varias desventajas, la principal de ellas se encuentra en los casos en los que no se tiene una transcripción a palabras de los audios de entrenamiento o que no se conoce la composición fonética de estas palabras, pero además, otra desventaja es que los resultados de este modelo son fuertemente dependientes de la cantidad de

estados K que se proponga para cada cadena de Markov y el número L de gaussianas con las que se modelan las emisiones de las mismas, teniendo que obtenerse estos valores de forma experimental [4]. Métodos de entrenamiento basados en inferencia variacional bayesiana [1, 5, 7] fueron propuestos para enfrentar esta última problemática, teniendo éstos la capacidad de aprender de forma automática la complejidad de la estructura, convergiendo a valores de K y L similares a los obtenidos experimentalmente [4, 6]. De forma similar, el problema de desconocer los fonemas del lenguaje puede modelarse ~~como~~ pensando que se tiene una cantidad infinita de unidades acústicas elementales, generadas mediante un proceso proceso de Dirichlet *stick-breaking* [9] en donde la probabilidad de cada una de estas unidades acústicas esta dada por

$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1}(1 - v_j)$$

con

$$v_i \sim \text{Beta}(1, \gamma)$$

Puede observarse en la forma en la que se construye la probabilidad de cada una de las unidades acústicas, que a medida que la cantidad de éstas aumento, la probabilidad de las subsecuentes unidades tenderá a cero, por la que el modelo solo le asignará una masa de probabilidad no despreciable a una cantidad finita de ellas, encontrando el número de unidades que mejor describe el lenguaje [8].

3.2. Estudios conexos

Áreas de conocimiento relacionadas:

- Procesamiento de señales
- Procesos estocásticos
- Teoría de detección y estimación
- Cálculo de variaciones
- Teoría de información y codificación

3.3. Planificación y alcance de la tesis

Son objetivos específicos de esta tesis:

- Analizar de forma exhaustiva la teoría detrás de la inferencia variacional bayesiana aplicada a procesos de Dirichlet en el marco del procesamiento del habla de cero recursos.
- Implementar un reconocedor automático del habla entrenada con palabras aisladas de un lenguaje de vocabulario reducido mediante inferencia variacional bayesiana en la que la cantidad de estados K y el número de gaussianas L del modelo HMM-GMM sea descubierto de forma automática.
- Implementar un reconocedor automático del habla que se entrene de forma totalmente no supervisada, descubriendo las unidades acústicas del lenguaje bajo estudio. Implementación sobre un lenguaje de vocabulario reducido, inglés y español rioplatense.

Para alcanzar estos objetivos, los pasos a seguir y sus duraciones estimadas son:

- Estudio de la señal de habla, forma biológica en la que ésta se produce y sus cualidades principales. Caracterización por medio de coeficientes cepstrales en las frecuencias de mel [3]. Duración estimada: 1 mes.
- Estudio de los métodos usados tradicionalmente en reconocimiento automático del habla, sus pros y contras. Implementación de los mismos para las bases de datos TIDIGITS, TIMIT y SALA1. Duración estimada: 3 meses.
- Estudio de la inferencia variacional bayesiana, sus propiedades como alternativa al entrenamiento clásico. Pruebas sobre datos sintéticos e implementación de un reconocedor entrenado con palabras usando TIDIGITS. Duración estimada: 3 meses
- Estudio de los procesos de Dirichlet, su impacto en reconocimiento automático del habla de cero recursos. Implementación de un

reconocedor entrenado con unidades acústicas descubiertas de forma automática en las bases de datos TIDIGITS, TIMIT y SALA1. Duración estimada: 4 meses

- Comparación de resultados y redacción del informe final. Duración estimada: 1 mes

Referencias

[1] *Pattern recognition and machile learning*, Christopher M. Bishop, Springer, 2006

[2] *Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains*, B. JUANG, AT&T Technical Journal, Vol. 64, U.S.A, 1985.

[3] *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, X. Huang, A. Acero, y H.-W. Hon, Prentice Hall, 2001.

[4] *Variational nonparametric Bayesian hidden Markov model*, N. Ding and O. Zhijian, Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE, pp. 2098-2101.

[5] *Variational Bayesian Inference for Hidden Markov Models With Multivariate Gaussian Output Distributions*, C. Gruhl and B. Sick, 2016 CoRR.

[6] *Variational Bayesian Analysis For Hidden Markov Models*, C. A. McGrory and D. M. Titterington, Australian & New Zealand Journal of Statistics, 2009, pp. 227-244.

[7] *Variational Bayesian Model Selection for Mixture Distributions*, A. Corduneanu and C. Bishop, Proceedings Eighth International Conference on Artificial Intelligence and Statistics, 2001, pp. 27-34.

[8] *Variational Inference for Acoustic Unit Discovery*, L. Ondel, L. Burget and J. Černocký, 5th Workshop on Spoken Language Technology for Under-resourced Languages, Procedia Computer Science 81, 2016, pp. 80-86.

[9] *Variational Inference for Dirichlet Process Mixtures*, D. Blei and M. Jordan, Bayesian Analysis, 2006, pp. 121-144.