Método para detección y seguimiento de objetos con aplicaciones en Realidad Aumentada

Christian Nicolás Pfarher
Director: *Dr. Albornoz, Enrique Marcelo*
Co-Director:*Dr. Martínez, Cesar*

FICH UNL sinc(*i*)

Centro de Investigación en señales
sistemas e **in**teligencia **c**omputacional

Ingeniería Informática - Universidad Nacional del Litoral

31 de Junio, 2013

Contenido

Contenido

Reconocimiento automático del habla

Reconocimiento automático del habla

Reconocimiento automático del habla

Modelo acústico y diccionario fonético:

Modelo de Lenguaje y Red de palabras:

Estructuración del habla y rasgos prosódicos

sinc(i)

Estructuración del habla y rasgos prosódicos

sinc(i)

sinc(i)

Rasgos prosódicos

El término prosodia refiere principalmente a tres características físicas del lenguaje hablado:

- Energía
- Frecuencia fundamental (F_0)*
- Duración del núcleo vocálico
- y otras ...

*La sensación auditiva de la F_0 es el tono de la voz o entonación.

Motivación: rasgos prosódicos

Comparación: **papá** y **capa**

Motivación: rasgos prosódicos

Comparación: **papá** y **capa**

Antecedentes

[Milone et al., 2003]: combinación de rasgos prosódicos (F_0, energía, duración temporal del núcleo vocálico, etc.) y la acentuación según las reglas ortográficas españolas.

[Chen et al., 2003]: HMM para modelar explícitamente la duración. Reetiquetando palabras y fonemas que estén al inicio y al final de frase generan un nuevo modelo de lenguaje.

[Szaszák and Vicsi, 2007]:. HMM entrenados con rasgos prosódicos, usado para segmentar unidades prosódicas. El sistema de RAH utiliza una red de palabras adicional ajustada en base a la salida del segmentador prosódico.

[Ananthakrishnan and Narayanan, 2007]: clasificador binario de acentuación silábico en base a la prosodia, luego las secuencias acentuales de una palabra son comparadas con las de la acentuación ToBI. Ésto en un RAH es estándar se usa para refinar las N-best.

[Huang and Renals, 2008]: se categorizan las características prosódicas a nivel de sílabas con cuantización vectorial y las palabras son concatenación de éstas. Estas definiciones componen un modelo de lenguaje adicional.

Objetivos

- Analizar los errores cometidos por un sistema de RAH estándar.
- Enfocar el estudio a los segmentos acústicos conflictivos para el sistema de RAH.
- Generar una nueva base de datos remuestreando el Corpus original.
- Diseñar un método para clasificar hipótesis verdaderas y falsas.
- Comprobar la capacidad del método de clasificación.

Contenido

Corpus y definición de un nuevo corpus

Base de datos

- Corpus Geográfico de la base de datos de habla Albayzin.
- Entrenamiento: 4400 elocuciones de 88 hablantes.
- Prueba: 2400 elocuciones de 48 hablantes.

Remuestreo del corpus

- Se utiliza el sistema de RAH entrenado.
- Se obtienen los N-best de cada frase y se segmenta por cada hipótesis.
- Si coincide con la etiqueta real, es una hipótesis verdadera.
- Si no coincide con la etiqueta real, es una hipótesis falsa.
- Se realizó un remuestreo para balancear el nuevo corpus.

Corpus y definición de un nuevo corpus

Base de datos

- Corpus Geográfico de la base de datos de habla Albayzin.
- Entrenamiento: 4400 elocuciones de 88 hablantes.
- Prueba: 2400 elocuciones de 48 hablantes.

Remuestreo del corpus

- Se utiliza el sistema de RAH entrenado.
- Se obtienen los N-best de cada frase y se segmenta por cada hipótesis.
- Si coincide con la etiqueta real, es una hipótesis verdadera.
- Si no coincide con la etiqueta real, es una hipótesis falsa.
- Se realizó un remuestreo para balancear el nuevo corpus.

Corpus y definición de un nuevo corpus

Base de datos

- Corpus Geográfico de la base de datos de habla Albayzin.
- Entrenamiento: 4400 elocuciones de 88 hablantes.
- Prueba: 2400 elocuciones de 48 hablantes.

Remuestreo del corpus

- Se utiliza el sistema de RAH entrenado.
- Se obtienen los N-best de cada frase y se segmenta por cada hipótesis.
- Si coincide con la etiqueta real, es una hipótesis verdadera.
- Si no coincide con la etiqueta real, es una hipótesis falsa.
- Se realizó un remuestreo para balancear el nuevo corpus.

Corpus y definición de un nuevo corpus

Clasificación de hipótesis

Características

- Parámetros prosódicos: F_0, Energía, F_1, ancho de banda de F_1, F_2 y ancho de banda de F_2.
- Los FV tienen: valores mínimos, medios, máximos, desviación estándar, asimetría y curtosis.
- Se incluyó: la distancia mínima y máxima entre F_1 y F_2, el ECM entre F_1 y F_2, y las pendientes de F_0, F_1 y F_2.

Clasificadores

- Vectores de características usando categorización según capacidad de discriminación (F-Score).
- Support Vector Machines. Validación cruzada con datos de entrenamiento.
- Entrenamiento y prueba del mejor modelo SVM por cada palabra.

Clasificación de hipótesis

Características

- Parámetros prosódicos: F_0, Energía, F_1, ancho de banda de F_1, F_2 y ancho de banda de F_2.
- Los FV tienen: valores mínimos, medios, máximos, desviación estándar, asimetría y curtosis.
- Se incluyó: la distancia mínima y máxima entre F_1 y F_2, el ECM entre F_1 y F_2, y las pendientes de F_0, F_1 y F_2.

Clasificadores

- Vectores de características usando categorización según capacidad de discriminación (F-Score).
- Support Vector Machines. Validación cruzada con datos de entrenamiento.
- Entrenamiento y prueba del mejor modelo SVM por cada palabra.

Contenido

Experimentos: extracción de características

- Se seleccionaron 12 de las palabras que más confunde el sistema de RAH.
- Los errores se calcularon en la etapa de extracción de N-Best.
- Se utilizó la biblioteca Praat para extraer los rasgos prosódicos.
- Para cada palabra, se genera una partición balanceada de entrenamiento y prueba (80 %-20 %).
- Se realizaron experimentos con datos crudos y normalizados[a].

[a] Cada dimensión se normalizó de forma independiente, utilizando su máximo y mínimo. Los factores de escala se usaron en la prueba.

Experimentos: extracción de características

- Se seleccionaron 12 de las palabras que más confunde el sistema de RAH.
- Los errores se calcularon en la etapa de extracción de N-Best.
- Se utilizó la biblioteca Praat para extraer los rasgos prosódicos.
- Para cada palabra, se genera una partición balanceada de entrenamiento y prueba (80 %-20 %).
- Se realizaron experimentos con datos crudos y normalizados[a].

[a] Cada dimensión se normalizó de forma independiente, utilizando su máximo y mínimo. Los factores de escala se usaron en la prueba.

Experimentos: extracción de características

- Se seleccionaron 12 de las palabras que más confunde el sistema de RAH.
- Los errores se calcularon en la etapa de extracción de N-Best.
- Se utilizó la biblioteca Praat para extraer los rasgos prosódicos.
- Para cada palabra, se genera una partición balanceada de entrenamiento y prueba (80 %-20 %).
- Se realizaron experimentos con datos crudos y normalizados[a].

[a] Cada dimensión se normalizó de forma independiente, utilizando su máximo y mínimo. Los factores de escala se usaron en la prueba.

Experimentos: extracción de características

- Se seleccionaron 12 de las palabras que más confunde el sistema de RAH.
- Los errores se calcularon en la etapa de extracción de N-Best.
- Se utilizó la biblioteca Praat para extraer los rasgos prosódicos.
- Para cada palabra, se genera una partición balanceada de entrenamiento y prueba (80 %-20 %).
- Se realizaron experimentos con datos crudos y normalizados[a].

[a] Cada dimensión se normalizó de forma independiente, utilizando su máximo y mínimo. Los factores de escala se usaron en la prueba.

F-Score: capacidad discriminativa

- Para cada palabra se mide el F-Score, para categorizar las características en base a su capacidad discriminativa.
- Dados los vectores de características FV_k, la puntuación se computa considerando las instancias Verdaderas (N_T) y las Falsas (N_F):

$$F(i) = \frac{\left(\bar{x}_i^{(T)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(F)} - \bar{x}_i\right)^2}{\frac{1}{N_T - 1}\sum_{j=1}^{N_T}\left(x_{j,i}^{(T)} - \bar{x}_i^{(T)}\right)^2 + \frac{1}{N_F - 1}\sum_{j=1}^{N_F}\left(x_{j,i}^{(F)} - \bar{x}_i^{(F)}\right)^2}$$

donde $\bar{x}_i$ es el promedio de la i-ésima característica, $\bar{x}_i{}^{(F)}$ y $\bar{x}_i{}^{(T)}$ son los promedios de las instancias Falsas y Verdaderas respectivamente, y $x_{j,i}$ es la i-ésima característica en la j-ésima instancia.

F-Score: capacidad discriminativa

- Para cada palabra se mide el F-Score, para categorizar las características en base a su capacidad discriminativa.
- Dados los vectores de características FV_k, la puntuación se computa considerando las instancias Verdaderas (N_T) y las Falsas (N_F):

$$F(i) = \frac{\left(\bar{x}_i^{(T)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(F)} - \bar{x}_i\right)^2}{\frac{1}{N_T - 1}\sum_{j=1}^{N_T}\left(x_{j,i}^{(T)} - \bar{x}_i^{(T)}\right)^2 + \frac{1}{N_F - 1}\sum_{j=1}^{N_F}\left(x_{j,i}^{(F)} - \bar{x}_i^{(F)}\right)^2}$$

donde $\bar{x}_i$ es el promedio de la i-ésima característica, $\bar{x}_i{}^{(F)}$ y $\bar{x}_i{}^{(T)}$ son los promedios de las instancias Falsas y Verdaderas respectivamente, y $x_{j,i}$ es la i-ésima característica en la j-ésima instancia.

Primer experimento

- Se crean 12 conjuntos de características en base a la categorización.
- Para cada conjunto se buscan los parámetros que definan el mejor modelo de clasificación.
- Los SVM utilizan un kernel de funciones de base radial y su precisión se mide usando un esquema de validación cruzada de 5 particiones.
- Finalmente, se obtienen los SVM definidos por los mejores parámetros para cada conjunto de características, usando datos de entrenamiento.
- Se utilizaron datos en crudo y normalizados.

Primer experimento

- Se crean 12 conjuntos de características en base a la categorización.
- Para cada conjunto se buscan los parámetros que definan el mejor modelo de clasificación.
- Los SVM utilizan un kernel de funciones de base radial y su precisión se mide usando un esquema de validación cruzada de 5 particiones.
- Finalmente, se obtienen los SVM definidos por los mejores parámetros para cada conjunto de características, usando datos de entrenamiento.
- Se utilizaron datos en crudo y normalizados.

Primer experimento

- Se crean 12 conjuntos de características en base a la categorización.
- Para cada conjunto se buscan los parámetros que definan el mejor modelo de clasificación.
- Los SVM utilizan un kernel de funciones de base radial y su precisión se mide usando un esquema de validación cruzada de 5 particiones.
- Finalmente, se obtienen los SVM definidos por los mejores parámetros para cada conjunto de características, usando datos de entrenamiento.
- Se utilizaron datos en crudo y normalizados.

Primer experimento

- Se crean 12 conjuntos de características en base a la categorización.
- Para cada conjunto se buscan los parámetros que definan el mejor modelo de clasificación.
- Los SVM utilizan un kernel de funciones de base radial y su precisión se mide usando un esquema de validación cruzada de 5 particiones.
- Finalmente, se obtienen los SVM definidos por los mejores parámetros para cada conjunto de características, usando datos de entrenamiento.
- Se utilizaron datos en crudo y normalizados.

Primer experimento

- Se crean 12 conjuntos de características en base a la categorización.
- Para cada conjunto se buscan los parámetros que definan el mejor modelo de clasificación.
- Los SVM utilizan un kernel de funciones de base radial y su precisión se mide usando un esquema de validación cruzada de 5 particiones.
- Finalmente, se obtienen los SVM definidos por los mejores parámetros para cada conjunto de características, usando datos de entrenamiento.
- Se utilizaron datos en crudo y normalizados.

Resultados para datos de entrenamiento en crudo

Word	*42*	*32*	*26*	*21*	*16*	*14*	*12*	*10*	*8*	*6*	*4*	*2*
CABO	61.11	63.89	63.89	63.89	63.89	63.89	70.83	76.39	77.78	**79.17**	76.39	62.50
CAUDAL	76.52	80.30	80.30	80.68	80.68	79.55	80.68	**85.61**	84.85	84.47	80.30	77.65
DESEMBOCA	80.38	80.38	80.38	80.38	80.38	80.38	80.38	80.38	80.38	**84.21**	80.38	74.64
DESEMBOCAN	79.79	79.79	79.79	83.94	84.72	84.72	84.72	84.72	84.72	**85.49**	**85.49**	63.73
MENOR	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	74.89
MENOS	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**
NOMBRE	86.75	86.75	87.73	87.83	**87.93**	87.73	87.54	87.63	87.63	86.95	86.75	79.49
NUMERO	84.85	84.85	84.85	84.85	84.85	84.85	84.85	84.85	84.85	84.85	**89.39**	87.88
PASA	79.17	79.17	80.11	80.11	80.30	80.11	79.36	79.36	79.36	79.55	**80.49**	73.48
PASAN	56.22	56.22	56.22	56.22	56.22	56.22	56.22	56.22	57.25	59.33	61.66	**65.80**
TIENE	52.66	52.66	52.66	52.66	52.66	52.66	52.66	**76.86**	75.27	73.27	71.94	65.82
TIENEN	69.08	69.08	69.08	71.60	72.27	72.10	**73.95**	73.45	73.61	70.92	68.57	64.71

Resultados para datos de entrenamiento normalizados

Word	*42*	*32*	*26*	*21*	*16*	*14*	*12*	*10*	*8*	*6*	*4*	*2*
CABO	77.78	81.94	77.78	81.94	86.11	86.11	**90.28**	73.61	75.00	75.00	75.00	66.67
CAUDAL	**89.77**	89.02	87.50	85.61	83.33	82.95	84.47	86.36	85.23	81.82	76.14	74.62
DESEMBOCA	84.93	**85.65**	85.41	82.78	84.21	82.78	83.49	81.34	81.82	71.05	65.79	61.24
DESEMBOCAN	80.83	80.05	**81.87**	79.53	79.02	78.50	77.72	75.13	78.24	69.95	62.69	58.55
MENOR	88.31	88.31	**89.18**	85.71	87.01	86.58	85.28	86.15	84.85	83.12	75.32	73.16
MENOS	86.39	**87.07**	85.71	86.39	86.39	85.71	85.71	85.03	82.99	84.35	82.31	73.47
NOMBRE	**88.32**	**88.32**	88.22	87.34	85.87	86.46	83.91	80.77	79.39	73.80	72.42	71.64
NUMERO	**89.39**	86.36	84.85	86.36	80.30	78.79	83.33	84.85	81.82	81.82	81.82	75.76
PASA	84.28	83.90	**84.47**	83.14	81.82	79.73	79.17	77.46	74.43	74.05	69.89	69.32
PASAN	74.61	**76.42**	74.35	75.13	72.80	74.09	75.13	74.61	72.54	68.13	69.43	65.54
TIENE	**78.19**	77.79	75.93	76.46	73.01	73.54	73.40	71.68	69.81	68.35	66.22	63.16
TIENEN	**75.97**	74.62	72.61	72.94	72.61	70.42	71.26	66.22	67.56	64.37	63.53	64.20

Segundo experimento

- Se entrenó un modelo SVM con todos los datos de entrenamiento para cada palabra, utilizando las configuraciones que obtuvieron mejores resultados en el *experimento 1*.
- Todos estos modelos SVM fueron probados con la partición de prueba.
- El promedio de clasificación fue de $77{,}66\,\%$ y $82{,}12\,\%$ utilizando datos crudos y normalizados respectivamente.

Segundo experimento

- Se entrenó un modelo SVM con todos los datos de entrenamiento para cada palabra, utilizando las configuraciones que obtuvieron mejores resultados en el *experimento 1*.
- Todos estos modelos SVM fueron probados con la partición de prueba.
- El promedio de clasificación fue de 77,66 % y 82,12 % utilizando datos crudos y normalizados respectivamente.

Segundo experimento

- Se entrenó un modelo SVM con todos los datos de entrenamiento para cada palabra, utilizando las configuraciones que obtuvieron mejores resultados en el *experimento 1*.
- Todos estos modelos SVM fueron probados con la partición de prueba.
- El promedio de clasificación fue de $77{,}66\,\%$ y $82{,}12\,\%$ utilizando datos crudos y normalizados respectivamente.

Resultados para datos de test con datos crudos

Palabra	Vector de características	Clasificación[%]
CABO	6	66.67
CAUDAL	10	74.24
DESEMBOCA	6	89.42
DESEMBOCAN	6	85.42
MENOR	42	77.19
MENOS	42	67.57
NOMBRE	16	90.20
NUMERO	4	75.00
PASA	4	81.06
PASAN	2	56.25
TIENE	10	82.98
TIENEN	12	85.91
Promedio		**77.66**

Resultados para datos de test con datos normalizados

Palabra	Vector de características	Clasificación[%]
CABO	12	66.67
CAUDAL	42	84.85
DESEMBOCA	32	89.42
DESEMBOCAN	26	82.29
MENOR	26	91.23
MENOS	32	83.78
NOMBRE	42	85.49
NUMERO	42	81.25
PASA	26	81.82
PASAN	32	77.08
TIENE	42	75.00
TIENEN	42	86.57
Promedio		**82.12**

Contenido

Conclusiones

- Se presentó un método dirigido a mejorar el rendimiento de un sistema de RAH, el cual considera las redes de palabras y la información prosódica.
- El método extrae las hipótesis de palabras de las redes de palabras generadas por un sistema de RAH estándar. Éstas representan las entradas de clasificadores de palabras que distinguen entre hipótesis verdaderas y falsas, utilizando información prosódica.
- Los resultados experimentales informan una tasa de clasificación promedio de 82 % sobre el conjunto de palabras seleccionado del corpus Albayzin.
- Este método puede ser aplicado en cualquier idioma, de similares características, ya que no incluye reglas específicas del español.

Conclusiones

- Se presentó un método dirigido a mejorar el rendimiento de un sistema de RAH, el cual considera las redes de palabras y la información prosódica.
- El método extrae las hipótesis de palabras de las redes de palabras generadas por un sistema de RAH estándar. Éstas representan las entradas de clasificadores de palabras que distinguen entre hipótesis verdaderas y falsas, utilizando información prosódica.
- Los resultados experimentales informan una tasa de clasificación promedio de 82 % sobre el conjunto de palabras seleccionado del corpus Albayzin.
- Este método puede ser aplicado en cualquier idioma, de similares características, ya que no incluye reglas específicas del español.

Conclusiones

- Se presentó un método dirigido a mejorar el rendimiento de un sistema de RAH, el cual considera las redes de palabras y la información prosódica.
- El método extrae las hipótesis de palabras de las redes de palabras generadas por un sistema de RAH estándar. Éstas representan las entradas de clasificadores de palabras que distinguen entre hipótesis verdaderas y falsas, utilizando información prosódica.
- Los resultados experimentales informan una tasa de clasificación promedio de 82 % sobre el conjunto de palabras seleccionado del corpus Albayzin.
- Este método puede ser aplicado en cualquier idioma, de similares características, ya que no incluye reglas específicas del español.

Conclusiones

- Se presentó un método dirigido a mejorar el rendimiento de un sistema de RAH, el cual considera las redes de palabras y la información prosódica.
- El método extrae las hipótesis de palabras de las redes de palabras generadas por un sistema de RAH estándar. Éstas representan las entradas de clasificadores de palabras que distinguen entre hipótesis verdaderas y falsas, utilizando información prosódica.
- Los resultados experimentales informan una tasa de clasificación promedio de 82 % sobre el conjunto de palabras seleccionado del corpus Albayzin.
- Este método puede ser aplicado en cualquier idioma, de similares características, ya que no incluye reglas específicas del español.

Trabajos futuros

- Integrar este método en un sistema de RAH estándar para incrementar las probabilidades de las hipótesis verdaderas en la red de palabras.

- Desarrollar un sistema de RAH que incluya este método en *una sola pasada*, de forma que se consideren los alineamientos de las hipótesis luego de su identificación acústica.

Trabajos futuros

- Integrar este método en un sistema de RAH estándar para incrementar las probabilidades de las hipótesis verdaderas en la red de palabras.
- Desarrollar un sistema de RAH que incluya este método en *una sola pasada*, de forma que se consideren los alineamientos de las hipótesis luego de su identificación acústica.