Clasificación de acento por región en hablantes estadounidenses

Lopez, Mendoza, Zacarías

***Resumen*— El problema de la identificación automática del acento es importante para varias aplicaciónes, como el perfil y el reconocimiento del hablante, así como también para mejorar los sistemas de reconocimiento automatico de voz (ASR). La naturaleza acentuada del habla se puede atribuir principalmente a la influencia de la cultura regional, la descendencia del hablante en la grabación de voz dada. En este trabajo proponemos una comparativa entre dos métodos para la clasificación de hablantes según su región.**

***Palabras clave*— Perceptrón multicapa(MLP), Red neuronal profunda(DNN), identificación de acento, Coeficientes Cepstrales de Mel (MFCC), extracción de características.**

I. Introducción

La capacidad de estimar y caracterizar el acento de un hablante proporcióna información valiosa en el desarrollo de sistemas de habla mas efectivos como reconocimiento de voz, etiquetado de flujo de audio en un documento hablado, monitoreo de canales o conversión de voz.

Grupos de personas con características geográficas, lingüísticas, sociales y antecedentes culturales similares entre otras cosas, comparten patrones comunes en su discurso, dando como resultado la impresión de un acento particular cuando hablan. Estos patrones de habla contribuyen al acento de una persona a partir de factores de: pronunciamientos y acústicos, uso de vocales particulares y sonidos de consonantes con sus variaciónes cuando se combinan en palabras y grupos de palabras, factores de presión, tempo, rítmicos y entonaciónales

Con la intención de elaborar un sistema capaz de identificar acentos entre distintos hablantes de idioma Ingles se elaboro la propuesta de construir dos modelos distintos que sean capaces de realizar dicha tarea.

La primer propuesta consta de un enfoque centrado en la extracción de características de las señales de audio de los distintos hablantes[1]. A partir de las señales de audio crudo de la base de datos realizamos un ventaneo a la señal donde para cada ventana del audio extraemos un conjunto de características que se consideran reelevantes a la hora del reconocimiento del acento en dicho audio, como ser los Cepstral Mel Coefficients (MFCC) y las Frecuencias Formantes. Con estos datos realizamos un vector de características extraídas de la ventana procesada y Habiendo obtenido las características de las ventanas de una señal procedemos a entrenar un Perceptron MultiCapa.

La segunda propuesta se basa en un enfoque sin extracción de características de la señal y un modelo de red neuronal profunda para la clasificación de acentos de los hablantes. A partir de la señal de audio nuevamente realizamos un ventaneo de las señales y utilizamos estas como patrones de entrada a la Red Neuronal Profunda, la cual se encargará de aprender a identificar el tipo de acento de cada ventana.

Ambos modelos seran entrenados y probados con la base de datos TIMIT que se compone de audios de hablantes nativos del idioma inglés.

II. Modelo de identificación del acento mediante MLP

El modelo de identificación del acento mediante el uso de un Perceptron Multicapa (MLP) sigue una secuencia de pasos (Fig.1) que se efectúan sobre la señal de audio. La red neuronal se entrena para clasificar vectores de características extraidas a partir de ventanas de los audios. Una vez que obtenemos el modelo entrenado, en la fase de prueba se efectuará para cada audio de la partición de Test, una clasificación de los vectores de características para las ventanas de ese audio. Habiendo pasado por todos los vectores, la clase que mas veces salió clasificada, será la clase que el modelo predice para el audio entero. A continuación se hara una descripción detallada de cada uno de los pasos que se siguieron para el desarrollo de este modelo.

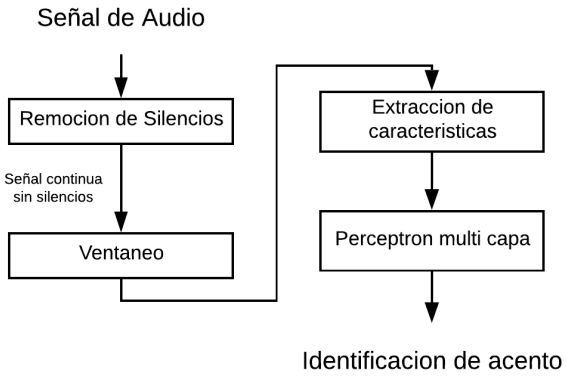


Fig.1 Diagrama de bloques del proceso de identificación de acentos para el enfoque con extracción de características

A. Eliminación de silencios

Utilizamos la señal de habla continua de entrada, eliminando las tramas en donde hay silencios. Una trama de silencio es aquella cuya energía es menor a 0.15 del promedio de la energía de toda la forma de la onda[1].

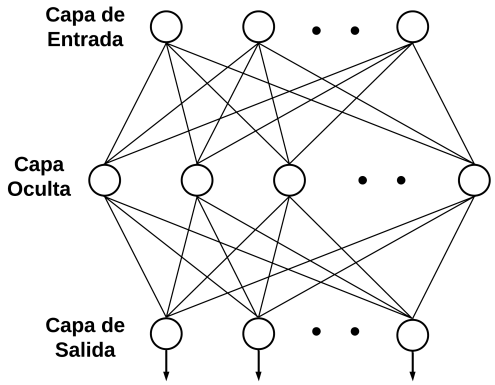
B. Ventaneo

En esta etapa se procesó la señal realizando un marco y ventaneo de la misma. Se utilizaron distintas longitudes de marco, desde los 20ms hasta los 200ms, con un solapamiento de entre 25% y 50% con el marco siguiente y anterior. Además dichos marcos fueron multiplicados por una ventana de Hamming para evitar el efecto de aliasing.

C. Extracción de características

Las dos primeras formantes (F1 y F2) permiten la identificación de las vocales, mientras que las siguientes tres (F3, F4 y F5) determinan el color de la voz. F3 está relaciónada con la acción de los labios, mientras que F4 y F5 varían con la anchura y longitud del tracto vocal [ref.formantes]. Los coeficientes cepstrales de frecuencia de mel (MFCC) se utilizan en muchas investigaciónes de identificación y clasificación de acento y proporciónan resultados aceptables[1]. En nuestro trabajo hemos tomado los trece primeros MFCC y las primeras 5 formantes. Las formantes fueron extraídas utilizando un filtro de predicción lineal (LPC), mientras que los MFCCs através de la transformada discreta de Fourier. Aquí además se realizó una compresión de las características realizando un promedio de cada 7 cuadros[1].

Fig.2 Diagrama de Perceptron Multicapa



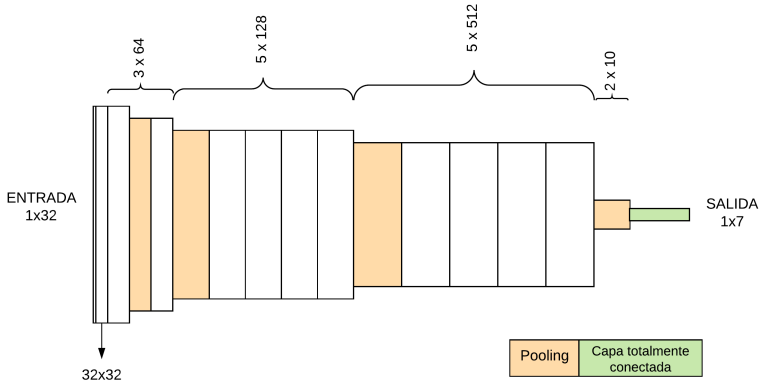
D. Propagación hacia adelante del perceptrón multicapa

Se utilizó un perceptrón multicapa de tres capas para clasificar las ventanas a las cuales se le extrajeron las características (Fig.2). El número de neuronas de la capa de entrada es igual al número de características extraídas, 18. La cantidad de neuronas en la capa oculta fué estipulada por prueba y error en realación al costo computaciónal y resultados. En la úlima capa el número de neuronas corresponde a la cantidad de regiones analizadas, en este caso 7. Las salidas de estas neuronas es un vector de numeros reales en el cual el índice del mayor elemento indica la clase ganadora. Se utilizó la función de unidad lineal rectificada por elementos (ReLU) como función de activación para la capa oculta y para la capa de salida, utilizamos la función SoftMax. En la etapa de entrenamiento del MLP se utilizó el método de aprendizaje de la entropía cruzada y un optimizador Adam. Además durante esta etapa se guarda el mejor modelo encontrado.

III. Modelo de identificación del acento mediante DNN

El modelo neuronal profundo consta de varias capas convoluciónales. Se probaron dos variantes. En la primera, se tienen capas convoluciónales de una dimensión para tratar los audios o ventanas como tales. En el segundo, usamos convoluciónes de dos dimensiones para tratarlos como si fueran imágenes. Estos modelos tienen la bondad de no necesitar extraer caracteríasticas de las señales si no que aprenden solas a realizar esta tarea a lo largo de toda su estructura. Ambos modelos, de una y dos dimensiones, son exactamente iguales en profundidad, sólo varía la cantidad de dimensiones que puede manejar. La estructura interna de estos siguen el siguiente patrón: se realiza una capa convoluciónal seguido de una ReLu, se realiza una normalización de Batch para evitar el desvanecimiento del gradiente y, en las capas 3, 5, 8 y 13, un MaxPooling. Al final, se realiza un AveragePool y seguidamente, una red totalmente conecetada con función de activación ReLu y salida SoftMax que cuenta con 7 neuronas. La función de pérdida utilizada es la entropía cruzada y se utiló también el optimizador Adam.

Diagrama modelo neuronal profundo



IV. Prueba y realización

A. Base de datos

Nuestra base de datos, TIMIT, contiene archivos de audio, con una frecuencia de muestreo de 16000Hz, obtenidos de diferentes personas con acentos distintos, distribuidos en hombres y mujeres de diferentes edades para cada acento. En todos los casos se enunciaron las mismas frases y palabras que involucran la mayoría de los sistemas. Los valores de las características fueron normalizados entre -1 y 1.
La distribución de los hablantes esta dada de la siguiente manera:

Región	Hombres	Mujeres	Total
1	31 (63%)	18 (27%)	49 (8%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
TOTAL	438 (70%)	192 (30%)	630 (100%)

Las regiones corresponden a New England, Northern, North Midland, South Midland, Southern, New York City y Western, respectivamente. La clase 8 fué descartada debido a que correspondía militares de todo el país, no aportando características a las clases principales.

Un aspecto importante sobre el uso de los datos fue la incorporación de tecnicas de Validación Cruzada a partir del metodo Hold Out en el modelo MLP, el cual se uso para dividir el set de datos de entrenamiento en una porción de entrenamiento del 80% de los datos y una porción de monitoreo del 20% restante. Este metodo resulto reelevante para la estimación de la capacidad de generalización del modelo durante la etapa de entrenamiento.
Para

En tanto al modelo DNN no se usaron tecnicas de validación cruzadas por el costo y tiempo que estas implicaban en el entrenamiento de dicho modelo.

B. Variación de los parámetros

Se tomaron distintos tamaño de ventana en ambos modelo, variando estos en ventanas de:

- 0.1s, solapadas en un 25%, es decir, con paso de 0.75s.
- 0.1s, solapadas en un 50%, es decir, con paso de 0.05s
- 0.1s, solapadas en un 75%, es decir, con paso de 0.25s.
- 0.050s, solapadas en un 50%, es decir, con paso de 0.025s

Además fué utilizada la ventana de Hamming para eliminar aliasing en todos los casos.
La cantidad de epocas se hizo variar entre las 1000 y las 7000, observando un estancamiento en la precision en alrededor de las 2000 epocas en los dos modelos.
Se probaron distintos tamaños de batch desde 1 a 2000, donde los resultados no variaron, pero en tamaños mas grandes, aumentó la velocidad de entrenamiento.
En cuanto a las características para el entrenamiento del MLP, se extrajo: la tasa de cruce cero, energía de la señal, entropía de la energía, centroide espectral, extensión espectral, entropía espectral, flujo espectral, desplazamiento espectral, los 13 primeros MFCC y las primeras 5 formantes. Se limitó solamente a los 13 primeros MFCC y las 5 formantes, debido a que las demas características aportaban ruido al sistema y elevaban el costo computaciónal.
En dicho modelo se variaron los parametros libres de la capa oculta, no teniendo insidencia en los resultados. Además se probó con una función de activación sigmoidea y una unidad lineal rectificada, obteniendo mejores resultados en esta útlima.
En el modelo neuronal profundo se probó con distintas convoluciónes. En el caso de convoluciónes 2D, se realizó un pre-procesamiento, entrenando con el espectrograma de mel (imagen) y ventaneo de audios, formando una matriz. Por el lado de convoluciónes 1D, se entrenó con la señal ventaneada y la señal en crudo. Se obtuvo una mejor eficiencia en la convolución 1D y mas especificamente cuando se entrenó con los audios en crudo.

V. Resultados

Para evaluar el desempeño de las redes se utilizó la entropía cruzada, comentado anteriormente. Para el MLP los resultados obtenidos fueron:

Tabla de resultados con distintas ventanas

MLP			
Ventana		Desempeño	
Tam Ventana	Solapamiento	Entrenamiento	Prueba
0,2 s	50.00 %	21.56 %	21.34 %
0,1 s	75.00 %	20.45 %	20.36 %
0,1 s	50.00 %	24.98 %	24.23 %
0,1 s	25.00 %	20.67 %	20.36 %
0,05 s	50.00 %	18.56 %	17.84 %

Donde las características utilizadas para el entrenamiento reportaron resultados similares en comparación al tamaño y solapamiento de ventana, por lo que solo se muestran los resultados con el más eficiente para evitar la redundancia.

MLP		
Características	Desempeño	
	Entrenamiento	Prueba
13 MFCC + 5 Formantes + Paquete energía espectral (*)	14.13 %	14.08 %
13 MFCC + 5 Formantes	24.98 %	24.23 %

Tabla de resultados según las características extraídas

Terminando con este método y con la misma lógica anterior la siguiente tabla muestra las pruebas de funciónes de activación

Tabla de resultados según la función de activación

MLP		
	Desempeño	
Función de activación	Entrenamiento	Prueba
Sigmoidea	20.96 %	20.21 %
Unidad lineal rectificada (ReLU)	24.98 %	24.23 %

Las pruebas del modelo DNN se hicieron sobre 2 principales variantes. La primera donde se entreno la red con la señal de audio entera sin modificaciónes y la segunda donde se procedio a ventanear la señal y utilizar las ventanas de

audio como patrones de entradas para la red profunda Ademas se realizaron pruebas pruebas mediante convoluciónes 1D. Los resultados del entrenamiento y prueba de estas variaciónes del modelo DNN se ven en la siguiente tabla:

Tabla de resultados del DNN

DNN			
		Desempeño	
Datos		**Entrenamiento**	**Prueba**
Sin procesar (en crudo)		**40,35%**	**32,12%**
Ventaneados			
Tam Ventana	**Solapamiento**		
0,2 s (1D)	50.00 %	24,43%	20,5%
0,1 s (1D)	75.00 %	22,54%	20,26%
0,1 s (1D)	50.00 %	22,89%	19,98%
0,1 s (1D)	25.00 %	23,78%	21,43%
0,05 s (1D)	50.00 %	19,88%	18,22%
1s (2D)	25.00%	19,79%	18,01%

VI. Conclusión

Se puede concluir que el modelo DNN obtiene mejor desempeño en comparación al MLP. El primero cuenta con un coste computaciónal mucho mayor debido a la cantidad de parametros de entrenamiento, pero esto se compensa con la ventaja de que los patrones de entrada no necesitan un procesamiento previo a diferencia del otro modelo. Para este ultimo, su mayor costo es intelectual, ya que se deben selecciónar características adecuadas para resolver el problema planteado, de esto dependerá su desempeño y eficiencia. En tanto a los aspectos que se pueden mejorar, se debería indagar mas sobre las características que pueden considerarse en la prosodia y sus métodos de extracción. Por otro lado, se podrían considerar aspectos que quedaron fuera del alcance de este trabajo, como redes neuronales dinámicas.

VII. Agradecimientos

Los autores de este documento quieren agradecer a Leandro Di Persia y toda la cátedra de Inteligencia Computaciónal de la Facultad de Ingeniería y Cs. Hídricas de la Universidad Naciónal del Litoral por la asistencia en el desarrollo del trabajo.

References

[1] Saeed Setayeshi Azam Rabiee. *Persian Accents Identification Using an Adaptive Neural Network*. Faculty of Nuclear Engineering and Physics Amirkabir University of Technology, 2010.