ESTUDIO DE LA INFLUENCIA DEL DESPLAZAMIENTO DE TRAMA EN PARAMETRIZACIÓN PARA SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DE HABLA

Javier Ordóñez Vázquez y Javier Macías Guarasa
Grupo de Tecnología del Habla. Dpto. Ingeniería Electrónica
Universidad Politécnica de Madrid
{jordonezv, macias} @die.upm.es - http://www-gth.die.upm.es

Palabras clave: parametrización, desplazamiento de trama variable

Resumen

En nuestro Grupo estamos trabajando en un amplio estudio y evaluación de las alternativas disponibles en la literatura al respecto de parametrización de habla con desplazamiento de trama variable. En este artículo se describen algunos experimentos iniciales en esta línea, que pretenden evaluar el efecto de utilizar distintos valores del desplazamiento de trama en parametrización con un amplio rango de condiciones de experimentación: distintos modelos acústicos, diferentes arquitecturas en cuanto al sistema de reconocimiento, etc., buscando propuestas de prometedoras variantes de la algorítmica al respecto.

1. Introducción

El primer módulo de cualquier sistema de reconocimiento automático de habla es un parametrizador que busca obtener las características más relevantes de la señal acústica para el reconocimiento. Para ello se hace un enventanado de la señal y se realizan los cálculos de parametrización con un intervalo determinado. Aquí se pretende evaluar el efecto de variar el intervalo con el que se calcula cada nueva trama (desplazamiento de trama) utilizando diferentes tipos de arquitecturas de reconocimiento (sistema no integrado e integrado) [3] y diferentes tipos de modelado (modelos discretos y semicontinuos, con dependencia e independencia del contexto). El objetivo que se pretende al variar el desplazamiento de ventana es uno de los siguientes:

- Buscar una mejora en la tasa de reconocimiento. Si se calculan tramas con un intervalo temporal menor es de esperar que se capturen con mayor fidelidad los cambios que se produzcan en la señal acústica y que las tasas de reconocimiento puedan mejorar [6].

- Buscar una disminución en la carga computacional del sistema de reconocimiento disminuyendo el número de tramas a procesar (aumentando el desplazamiento de trama) intentando que las tasas de reconocimiento bajen lo menos posible.

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

Típicamente se pueden utilizar dos aproximaciones a la hora de realizar los análisis con diferentes desplazamientos de ventana:

- Desplazamiento de ventana fijo: La separación en el cálculo entre tramas es constante.
- Desplazamiento de ventana variable: La separación entre las distintas ventanas de cálculo no es constante, calculándose en ciertas zonas tramas con mayor frecuencia que en otras.

Cada una de ellas presenta ciertas ventajas e inconvenientes y es nuestro objetivo en este artículo incluir algunos de los trabajos iniciales que forman parte de un estudio más amplio en el que se está trabajando en nuestro grupo, en la línea de revisar en detalle todas las alternativas disponibles en la literatura, diseñando nuevas estrategias al respecto y trabajando sobre un amplio conjunto de tareas, tipos de modelado y arquitecturas de reconocimiento.

2. Análisis con desplazamiento de ventana variable

2.1. Motivación

Los cambios que se producen en la señal acústica no son de ninguna manera uniformes a lo largo del tiempo: existen zonas de cambios rápidos (como por ejemplo las zonas de fonemas oclusivos) y zonas en las que existe una gran estacionariedad (zonas de silencios, vocales). No parece razonable entonces calcular las tramas de forma uniforme. Intuitivamente la aproximación más interesante sería calcular más tramas en aquellas zonas en las que la señal varíe de una forma más rápida y disminuir el número de tramas en las zonas de cambio lento, ya que las tramas cercanas temporalmente serán muy parecidas y no aportarán nueva información, salvo la temporal. En la bibliografía se han encontrado diferentes propuestas para realizar un análisis con desplazamiento de trama variable [6], [2], [4] y en todos ellos se parte de una parametrización realizada de forma uniforme (con desplazamiento de ventana fijo) y utilizando alguna medida de distancia, se eliminan las tramas que difieran de otras ya seleccionadas y que por tanto aporten poca información. Se han manejado fundamentalmente dos tipos de medidas de distancias:

- Distancia euclídea: Calculada bien entre tramas consecutivas [6] o bien entre la trama en curso y la última que no fue descartada [4].
- Norma de la derivada: Se calcula para cada trama la norma de la primera derivada (delta) [2], lo que presenta la ventaja de que en la mayor parte de los parametrizadores actuales se calcula dicho valor de forma estándar, de modo que parte del cálculo ya está realizado y, además, a diferencia de la distancia, el cálculo engloba a varias tramas, por lo que presentará una mayor inmunidad en caso de presencia de ruido.

Se ha realizado un estudio exhautivo de los diferentes métodos encontrados en la bibliografía, con el objetivo de comparar sus prestaciones en diferentes escenarios de experimentación: diferentes arquitecturas, modelos, etc. Aquí, como muestra del funcionamiento y posibilidades de los algoritmos de análisis con desplazamiento de trama variable, se presentarán los resultados obtenidos con dos de ello que usan las medidas de distancia descritas anteriormente.

2.2. Método de selección de tramas con distancia euclídea

En [6] se expone un método basado en la medida ponderada de la distancia eucídea entre tramas consecutivas. Allí se parte de una parametrización con

2,5 ms de desplazamiento de trama, con el objetivo de capturar los cambios más rápidos de la señal. La distancia se pondera por el logaritmo de la energía de la trama en curso, a la que se le sustrae la energía media de la pronunciación. En nuestro caso se va a partir de una parametrización con 10 ms de desplazamiento, con el objetivo de evaluar la bondad del método para eliminar todas las tramas posibles sin que varíe significativamente la tasa de reconocimiento, al tiempo que se realizan variaciones en la ponderación. El algoritmo básico es:

1. Se calcula la distancia euclídea entre tramas consecutivas

2. Se ponderan las distancias calculadas, usando en este caso el logaritmo de la energía de la trama (dando así mayor importancia a las tramas con más energía puesto que son más significativas en el sentido de que están menos corrompidas por el ruido) y restándole una constante. El valor de dicha constante, en lugar del logaritmo de la energía media de la palabra como se hacía en [6], es en nuestro caso el logaritmo de una estimación de la energía media del ruido calculada en el periodo de silencio antes de que empiece la locución. El objetivo de esta resta es penalizar aquellas tramas en las que su contenido sea básicamente ruido, ya que el valor de ponderación será entonces muy bajo. La ventaja de utilizar el silencio previo a la locución en lugar de la energía media, es que la ponderación se puede hacer de forma síncrona en trama.

3. Se calcula un umbral para cada locución dependiente del número de tramas que se pretendan eliminar. En nuestro caso se ha recurrido a un algoritmo iterativo de cálculo de umbral para controlar en todo momento la cantidad de tramas eliminadas.

4. Finalmente se procede a la selección de tramas: Se acumulan distancias desde el principio de la locución y el valor acumulado se compara con un umbral. En el caso de que se sobrepase el mismo, se acepta la trama (no se rechaza) y en ese momento se inicializa de nuevo el acumulador de distancias.

2.3. Método de selección de tramas con derivada

En [2] se propone un método basado en la norma de la derivada. En ese artículo no se acumulan distancias en el algoritmo de selección ni tampoco se realiza ponderación de tramas (en nuestros experimentos no logramos resultados positivos en las condiciones descritas). Esto es debido a que al no acumular distancias, dado un umbral, todas las tramas que estén por debajo de ese umbral serán rechazadas. El problema es que las zonas con norma más baja son las de silencios y vocales y al eliminar la información de estas últimas la tasa de reconocimiento baja sustancialmente.

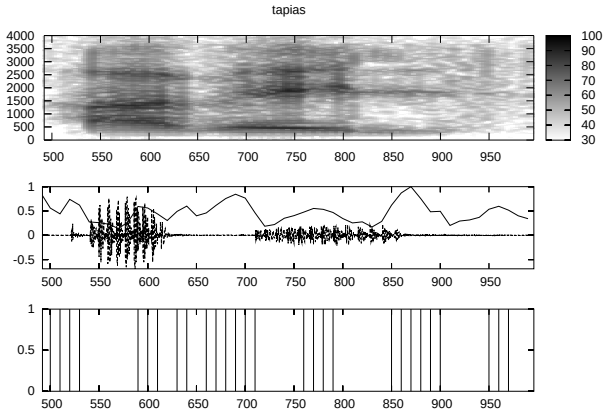

Figura 1: Selección de tramas con derivada sin acumular normas y sin ponderar

En la figura 1 se muestra el efecto de utilizar la selección variable de tramas descrita en [2] sobre la palabra "tapias". En la primera fila se muestra el espectrograma, en la segunda la norma de la derivada y en la tercera se marca el inicio de las tramas escogidas sobre una parametrización de 10 ms, en la que se descarta tramas hasta que el número es equivalente al de una parametrización fija de 17,5 ms (que es la que se utiliza posteriormente en los experimentos). Se puede observar cómo al no acumular distancias, en las zonas de norma pequeña no se selecciona ninguna trama, lo que repercute negativamente en la capacidad de discriminación del sistema. Por otra parte, al no usar ponderación, se escogen tramas en las zonas de silencio, en las que predomina el ruido.

En nuestro caso hemos optado por hacer uso de la idea de la norma de la derivada pero utilizando el algoritmo descrito en el apartado anterior, sin más que sustituir las distancias por normas.

En la figura 2 se muestra la selección de tramas efectuada por el método descrito en este artículo y que se evaluará posteriormente.

Como se puede observar, las tramas se escogen de una forma más uniforme que en el caso anterior (gracias a la acumulación de las normas), tomando más tramas en aquellas zonas en las que existe mayor variación. Se puede comprobar la importancia de la ponderación, que hace que en las zonas en las que predomina el ruido apenas se seleccionen tramas.

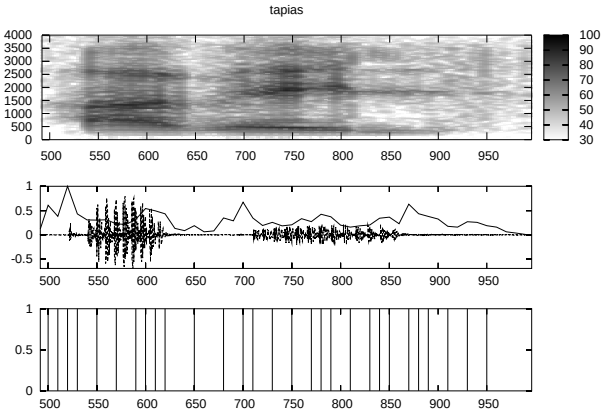

Figura 2: Selección de tramas con derivada con acumulación de normas y ponderando

3. Experimentos y resultados

Como muestra de los resultados de los experimentos realizados y para ilustrar lo dicho anteriormente, se ofrecerán inicialmente los resultados que se han obtenido utilizando diferentes desplazamientos fijos de ventana con cada uno de los sistemas bajo estudio. Además se mostrarán los resultados de utilizar análisis con desplazamiento variable de ventana para alguno de los casos.

3.1. Condiciones generales de experimentación

Los experimentos se han realizado utilizando la base de datos VESTEL [5] que contiene un total de 9.729 pronunciaciones con un tamaño de diccionario de 1946 palabras. Se ha utilizado 10 grupos de entrenamiento y otros tantos de reconocimiento utilizando la técnica de *leave-one-out* para mejorar la fiabilidad estadística de los resultados. En parametrización se han utilizado 10 parámetros MFCC más energía y su primera derivada. El reconocedor está basado en alófonos (45 unidades) y se utilizan modelos de Markov de tres estados cada uno, con autosaltos, saltos simple y dobles, tal como se hace en [1]. Los resultados se ofrecen teniendo en cuenta la fiabilidad estadística, ofreciendo la banda en la que se encuentra el resultado con una fiabilidad del 95 %.

3.2. Experimentos con desplazamiento de ventana fijo

Nuestro objetivo inicial era evaluar distintos valores de desplazamiento haciendo uso de los siguientes marcos de experimentación:

- Arquitectura no integrada: Se evaluarán las prestaciones en un sistema de bajo rendimiento diseñado para servir como sistema de hipótesis en una arquitectura hipótesis-verificación [3]. Está formado por un generador de cadena fonética seguido por un módulo de acceso léxico. Se utilizarán tanto modelos discretos como semicontinuos, independientes del contexto.
- Arquitectura integrada (se hace uso de la información acústica y de las limitaciones impuestas por el léxico de forma conjunta). Se evaluará con unidades independientes del contexto (modelos discretos y semicontinuos) y dependientes del contexto (modelos semicontinuos).

Modelos	Desplazamiento de ventana		
	2.5 ms	10 ms	17.5 ms
discreto	40.71 %±0.98	45.78 %±0.99	44.11 %±0.99
semicont.	49.93 %±0.99	57.51 %±0.98	57.47 %±0.98

Cuadro 1: Resultados sistema no integrado

Modelos	Desplazamiento de ventana		
	2.5 ms	10 ms	17.5 ms
discreto	53.68 %±0.99	66.63 %±0.94	65.24 %±0.95
semicont.	61.50 %±0.97	76.15 %±0.85	77.06 %±0.84

Cuadro 2: Resultados sistema integrado (modelos independientes del contexto)

A la vista de los resultados obtenidos, pueden obtenerse las siguientes conclusiones:

1. En contra de lo que en un primer momento pudiera pensarse, aumentar el número de tramas no tiene por qué llevar necesariamente a mejores resultados en la tasa de reconcimiento. Para todos los experimentos llevados a cabo se produce una bajada significativa en la tasa de reconocimiento en el caso de utilizar el desplazamiento de trama de 2,5 ms. En [2] se da justifica dicho efecto dado que un número demasiado elevado de tramas en relación al número de estados de los HMM's puede provocar que aumente el número de inserciones, con la consiguiente bajada en la tasa de reconocimiento.

Modelos	Desplazamiento de ventana		
	2.5 ms	10 ms	17.5 ms
semicont	73.74 %±0.92	84.92 %±0.75	86.28 %±0.72

Cuadro 3: Resultados sistema integrado (modelos dependientes del contexto)

2. Para todas las arquitecturas, en el caso de utilizar modelos semicontinuos, la tasa es equiparable e incluso mejora si reducimos el número de tramas (aumentando el desplazamiento de trama hasta a 17,5 ms). Esto implica disminuir en un 42.86 % el número de tramas, con lo que ello conlleva en cuanto a ahorro computacional.
3. Al usar modelos discretos es más acusado el efecto de la eliminación de tramas que en el caso de utilizar los semicontinuos.

3.3. Experimentos con desplazamientos de ventana variables

Vistos los resultados del experimento anterior, se puede comprobar cómo en el caso de la utilización de modelos discretos con la arquitectura no integrada, el descenso relativo de la tasa por eliminar tramas (17,5 ms) con respecto a la parametrización con 10 ms es de un 3,65 %. En este apartado aplicaremos los métodos de selección de tramas explicados anteriormente para partiendo de la parametrización de 10

ms descartar tramas de forma no uniforme hasta que la parametrización sea equivalente a una de 17,5 ms. Los resultados se compararán con los obtenidos con las parametrizaciones a 10 y 17,5 ms.

	Discretos		
	Resultado	(17.5 ms)	(10 ms)
Método distancia	45.63 %±0.99	3.45 %	-0.33 %
Método derivada	46.75 %±0.99	5.97 %	2.12 %

Cuadro 4: Resultados para los métodos con desplazamiento de trama variable

En este caso el análisis con desplazamiento de trama variable permite eliminar tramas (llegando a un desplazamiento equivalente de ventana de 17,5 ms) con tasas superiores a las obtenidas con la parametrización fija a 17,5 ms y equiparable a la parametrización con 10 ms de desplazamiento entre tramas.

4. Conclusión

En este artículo se ha evaluado el efecto de aumentar el desplazamiento fijo entre tramas o disminuirlo para diferentes arquitecturas: no integrada e integrada y con unidades independientes del contexto (modelos discretos y modelos semicontinuos) y dependientes del contexto (modelos semicontinuos). Se ha comprobado que aumentar el número de tramas calculadas (desplazamiento entre tramas de 2,5 ms) conduce en todos los casos a una degradación de la tasa. En cuanto al efecto de disminuir el número de tramas (desplazamiento entre tramas de 17,5 ms) se ha comprobado cómo en el caso de los modelos semicontinuos, la tasa no se ha visto perjudicada, con lo que es posible reducir la carga computacional del sistema. Se ha ensayado 2 técnicas de análisis con desplazamiento de trama variable para el caso de eliminar tramas utilizando un sistema no integrado y modelos semicontinuos, en el que con la estrategia de desplazamiento fijo, la tasa se veía perjudicada. Se ha comprobado cómo estas técnicas, para el mismo número de tramas (parametrización con desplazamiento de 17,5 ms) superan a la correspondiente parametrización con desplazamiento fijo y pueden igualar a la parametrización realizada con 10 ms.

Se plantea como línea futura de investigación el estudio de la relación entre el desplazamiento de ventana (o lo que es equivalente, número de tramas consideradas) y el número de estados de los HMM's para comprobar si es posible la obtención de buenos resultados con desplazamientos de trama bajos.

Igualmente estamos estudiando el efecto del uso de estos métodos cuando se combinan con técnicas de parametrización tipo Rasta, en el que hay una estrecha relación entre el desplazamiento de ventana usado y los parámetros de control del parametrizador, siendo éste sumamente sensible a variaciones en dicho desplazamiento, lo que dificulta la aplicabilidad general del método.

Además de las aplicaciones ya mencionadas para el análisis con desplazamiento de trama variable (capturar con mayor precisión los cambios rápidos en la señal y la reducción de la carga computacional) se constituyen como una buena opción para aquellos casos en los que la velocidad de los locutores sea heterogénea. En los casos en los que la velocidad de locución fuera elevado (a mayor velocidad, más diferencia entre tramas consecutivas) sería posible utilizar un desplazamiento de trama más pequeño o bien utilizar un desplazamiento de trama mayor en el caso de una velocidad de locución lenta.

Referencias

[1] R. Bakis. Continuous speech word recognition via centisecond acoustic states. *Proceedings of the Acoustical Society of America*, 1982.

[2] Philippe Le Cerf and Dirk Van Compernolle. A new variable frame rate analysis method for speech recognition. *IEEE Signal Processing Letters*, 1(12), 1994.

[3] Javier Macías Guarasa. *Arquitecturas y métodos en sistemas de reconocimiento automático de habla de gran vocabulario*. Tesis doctoral, ETSIT UPM, 2001.

[4] K.M. Ponting and S.M. Peeling. The use of variable frame rate in anlysis in speech recognition. *Computer Speech and Language*, 5(2):169–179, April 1991.

[5] D. Tapias, A. Acero, J. Esteve, and J.C. Torrecilla. The vestel telephone speech database. *ICSLP 94*, page 1811, 1994.

[6] Qifeng Zhu and Abeer Alwan. On the use of variable frame rate analysis in speech recognition. *ICASSP*, pages 3264–3267, 2000.