Anexo B: Conjuntos de características

1. IS09 INTERSPEECH 2009 Emotion Challenge

Este conjunto de características fue el primero de su clase, creado para el reconocimiento de la emociones del habla, basado en el trabajo de Schuller en (Schuller, Rigoll, & Lang, 2003), ya que diseñaron un conjunto de características minimalistas para modelos estáticos con descriptores acústicos de bajo nivel (LLDs) muy utilizados y conocidos en especial los coeficientes cepstrales (MFCCs). Este conjunto de características cuenta con 16 LLD (Tabla 3.1) de los cuales 12 son los primeros MFCC debido a su buen reconocimiento de emociones y 16 coeficientes de regresión delta. A estas 32 características se le aplican 12 funciones estadísticas (Tabla 3.2) que dan un total de 384 características para este conjunto.

Descriptor LLD	Cantidad
MFCC 1-12	12
F0 vía ACF	1
Raíz cuadrada media (RMS) de Energía	1
Tasa de cruce por cero (ZCR)	1
Probabilidad de la voz	1
Total de descriptores	16

Tabla 3. 1 Descriptores de bajo nivel IS09

Funciones	Cantidad
Promedio aritmético	1
Momentos: desviación estándar, asimetría, curtosis	3
Valores máximos	2
Valores mínimos	2
Rango	1
Regresión lineal	2
Error de regresión	1
Total de funciones	12

Tabla 3. 2 Funciones estadísticas del IS09

2. IS10 INTERSPEECH 2010 Paralinguistic Challenge

El objetivo del INTERSPEECH de 2010 era el reconocimientos de la edad y el nivel de interés (Schuller, Björn and Steidl, Stefan and Batliner, Anton and Burkhardt, Felix and Devillers, Laurence and Müller, Christian and Narayanan, 2013), debido a que esta tarea era nueva con respecto al desafío anterior el número de características aumento y se actualizaron los que ya estaban para que los participantes escogieran las características más importantes que cumplieran con el reto. En comparación con el IS09 se eliminó la tasa de cruce por cero (ZCR), se añadió el algoritmo de suma subarmónica en lugar de la función de auto correlación (ACF), además se adicionaron bandas logarítmicas de frecuencia melódica (MFB), frecuencias espectrales de línea (LSF) y descriptores de calidad de voz como el jitter y shimmer. Este conjunto contienen 38 LLDs y 38 coeficientes de regresión delta, todas estas se dividen en dos grupos (Tabla 3.3) el grupo A consta de 34 LLDs y sus respectivos coeficientes de regresión delta y el grupo B son los 4 LLDs y coeficientes delta restantes, al grupo A se le aplican 21 funciones (Tabla 3.4) dando como resultados 1428 características y al grupo B se le aplican 19 funciones (Tabla 3.4) que dan otras 152 características. Además se incluyen el número de onsets de la frecuencia fundamental y la duración en segundos del segmento esto para reflejar la velocidad del habla y la longitud de segmento respectivamente. Al final este conjunto el conjunto IS10 contiene 1582 características. Cabe resaltar que este conjunto descrito anteriormente se utilizó para el desafío de reconocimiento de intereses mientras que para el reconocimiento de edad se usó una versión reducida de este conjunto debido a la cantidad elevada de casos.

Descriptor LLD	**Cantidad**
Grupo A (34)	
Loudness	1
MFCC 0-14	15
Banda log de frecuencia melódica (MFB 1-8)	8
Pares de línea espectral 0.- 7(LSP)	8
Suma subarmonica	1
Probabilidad de voz	1
Grupo B (4)	
Frecuencia fundamental	1
Jitter	1
Delta jitter	1
Shimmer	1
Total de descriptores	38

Tabla 3. 3. Descriptores de bajo nivel para IS10

Funciones	Cantidad
Promedio aritmético	1
Media cuadrática	1
Momentos: desviación estándar, asimetría, curtosis	3
Cuartiles	6
99% Valor máximo	1
1 % Valor mínimo	1
Rango 1-99%	1
Regresión lineal	1
Error de regresión	1
Regresión cuadrática	2
Error de regresión cuadrática	1
Centroide	1
Media pico	2
Pico de distancia	2
Tamaño de segmento : Media, mínimo, máximo, desviación estándar	4
Tiempo de actualización	1
Nivel inferior	1
Tiempo de subida y bajada	2
Tiempo de curvatura	2
Ganancia LP	1
Coeficientes LP 1-5	5
Total de funciones	40

Tabla 3. 4 Funciones estadísticas para IS10

3.1.1 INTERSPEECH 2011 Speaker State Challenge Set

El reto de este INTERSPEECH era el reconocimiento del estado de la persona que se encuentra hablando, como por ejemplo si tiene somnolencia o intoxicación alcohólica, así que para esto se propuso el IS11 un conjunto con muchas más características y más completo que el anterior (Schuller, Björn and Steidl, Stefan and Batliner, Anton and Schiel, Florian and Krajewski, Jarek and Weninger, Felix and Eyben, 2014). Al igual que el IS10 este conjunto cuenta con dos grupos el grupo A con 54 descriptores entre energía, espectrales y cepstrales (Tabla 3.5) y el grupo B con 5 LLDs de fuente y señal de excitación. Para el grupo A se aplican 37 funciones mientras que para el grupo B son 36 dando como resultado en total 4356 características, además de los LLDs de fuerza bruta el IS11 tiene como plus que contiene descriptores temporales de segmentos con y sin voz que se obtienen aplicando otras funciones a la F0 y sus coeficientes delta de primer orden (valor mínimo y máximo, la media, la desviación estándar cuando la F0 >0), además como el IS10 la duración también hace parte del conjunto, así que se añaden otras 11 características es decir que en total son 4367 características las que conforman este conjunto.

Descriptor LLD	**Cantidad**
Grupo A (54)	
Loudness	1
Loudness modulado	1
Energía RMS	1
ZCR	1
Bandas RASTA 1-26	26
MFCC 0-12	12
Banda de energía	2
Puntos espectrales de Transferencia (RoP) : 25,50,75 Y 90%	4
Flujo espectral	1
Entropía espectral	1
Momentos espectrales : varianza espectral, asimetría espectral y curtosis espectral	3
Pendiente espectral	
Grupo B (5)	
Frecuencia fundamental	1
Jitter	1
Delta jitter	1
Shimmer	1
Probabilidad de voz	1
Total de descriptores	59

Tabla 3. 5. Descriptores de bajo nivel para IS11

Funciones	Cantidad
Promedio aritmético	1
Media cuadrática	1
Momentos: desviación estándar, asimetría, curtosis	3
Cuartiles	6
99% Valor máximo	1
1 % Valor mínimo	1
Rango 1-99%	1
Regresión lineal	1
Error de regresión	1
Regresión cuadrática	2
Error de regresión cuadrática	1
Centroide	1
Media pico	2
Pico de distancia	2
Tamaño de segmento : Media, mínimo, máximo, desviación estándar	4
Tiempo de actualización	1
Nivel inferior	1
Tiempo de subida y bajada	2
Tiempo de curvatura	2
Ganancia LP	1
Coeficientes LP 1-5	5
Total de funciones	40

Tabla 3. 5 Funciones estadísticas para IS11

3.1.2 INTERSPEECH 2012 Speaker Trait Challenge Set

Para este reto propone aún más características que los anteriores, pero reduce el hecho de aplicarle a todos los descriptores todas las funciones ya que cuando se hacen cálculos, estos no aportan una información significativa y es mejor eliminarlos. Al igual que los otros IS el IS12 se divide en 2 grupos de descriptores el grupo A con 58 LLDs entre energía, espectrales y cepstrales y el grupo B con 6 descriptores de fuente y señal de excitación (Tabla 3.6). A los 58 descriptores del grupo A se les aplican 58 funciones mientras que a sus deltas solo 38 y al grupo B se le aplican 56 funciones y a sus respectivos deltas solo 36 (Tabla3.7), como resultado se obtienen 6120 características además que cuenta con algunos descriptores temporales al igual que IS11. A diferencia del conjunto de IS11 la duración total del segmento no se tiene en cuenta como característica ya que según el objetivo del desafío que es el reconocimiento del rasgo del hablante, este fenómeno no es a corto plazo por lo que la duración de una sola expresión no diría nada en concreto con respecto al hablante. Finalmente este conjunto tiene un total de 6125 características.

El radio de valores no Zero (1) y las longitudes (4) de los segmentos donde la F0>0 son parte de los descriptores (valor mínimo y máximo, media y desviación estándar)

Descriptor LLD	**Cantidad**
Grupo A (58)	
Loudness	1
Loudnes modulado	1
Energía RMS	1
ZCR	1
Bandas RASTA 1-26	26
MFCC 0-14	14
Banda de energía	2
Puntos espectrales de Transferencia (RoP) : 25,50,75 Y 90%	4
Flujo espectral	1
Entropía espectral	1
Momentos espectrales : varianza espectral, asimetría espectral y curtosis espectral	3
Armonías	1
Pendiente espectral	1
Nitidez espectral	1
Grupo B (6)	
Frecuencia fundamental	1
Jitter	1
Delta jitter	1
Shimmer	1
Probabilidad de voz	1
Armónicos logarítmicos en relación al ruido (HNR)	1
Total de descriptores	65

Tabla 3. 6. Descriptores de bajo nivel para IS12

Funciones	Cantidad
Media aritmética	1
Media aritmética positiva	1
Media de la raíz cuadrática	1
Plenitud	1
Momentos: desviación estándar, asimetría, curtosis	3
Cuartiles	6
99% Valor máximo	1
1 % Valor mínimo	1
Rango 1-99%	1
Posición máxima y mínima	2
Rango full	1
Regresión lineal	1
Error de regresión	1
Regresión cuadrática	2
Error de regresión cuadrática	1
Centroide	1
Media pico	3
Pico de distancia	2
Rango de pico	2
Rango de valle	1
Pendientes pico-valles	4
Tamaño de segmento : Media, mínimo, máximo, desviación estándar	4
Tiempo de actualización	4
Nivel inferior	4
Tiempo de subida y bajada	2
Tiempo de curvatura	2
Ganancia LP	1
Coeficientes LP 1-5	5
Total de funciones	61

Tabla 3. 7 Funciones estadísticas para IS12

3.1.3 INTERSPEECH 2013 ComParE Set y INTERSPEECH 2014 ComParE Set

Para el conjunto de características ComParE, el IS12 ha aumentado en alrededor de 200 características. Se eliminaron las irregularidades y las redundancias, se han ajustado los parámetros y se han mejorado algunos algoritmos de extracción, al igual que los anteriores INTERSPEECH, el ComParE tiene dos grupos, el grupo A con 59 LLDs se le aplican 54 funciones y a sus respectivos 59 LLDs delta se aplican 46 funciones (Tabla 3.8) y esta el grupo B con 6 LLDs a las cuales se le aplican 39 funciones y a sus respectivos delta también se le aplican 39 funciones (Tabla 3.9), dando en total de 6368 características, además este conjunto contiene las mismas 5 estadísticas globales temporales que el IS12 y consta de aplicar a la F0 las siguientes funciones : relación de valores diferentes a 0, estadísticas de longitud (valor mínimo, máximo, media y desviación donde F0>0) es decir que el conjunto cuenta en total con máximo 6373 características.

El radio de valores no Zero (1) y las longitudes (4) de los segmentos donde la F0>0 son parte de los descriptores (valor mínimo y máximo, media y desviación estándar)

La segunda versión del ComParE se dio en el INTERSPEECH 2014 y no se tuvo la necesidad de modificar el ComParE de 2013 debido a su excelente rendimiento en tareas de procesamiento de habla y música, como ejemplo esta (Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013) por lo tanto el ComParE 2014 se usa para la clasificación de los niveles de carga respectivos de las grabaciones de voz

Descriptor LLD	**Cantidad**
Grupo A (58)	
Loudness	1
Loudnes modulado	1
Energía RMS	1
ZCR	1
Bandas RASTA 1-26	26
MFCC 1-14	14
Banda de energía	2
Puntos espectrales de Transferencia (RoP) : 25,50,75 Y 90%	4
Flujo espectral	1
Centroide espectral	1
Entropía espectral	1
Momentos espectrales : varianza espectral, asimetría espectral y curtosis espectral	3
Armonía	1
Pendiente espectral	1
Nitidez espectral	1
Grupo B (6)	
Frecuencia fundamental	1
Jitter	1
Delta jitter	1
Shimmer	1
Probabilidad de voz	1
Armónicos logarítmicos en relación al ruido (HNR)	1
Total de descriptores	66

Tabla 3. 8 Descriptores de bajo nivel para el ComParE

Funciones	**Cantidad**
Media aritmética	1
Media aritmética positiva	1
Media de la raíz cuadrática	1
Plenitud	1
Momentos: desviación estándar, asimetría, curtosis	3
Cuartiles	6
99% Valor máximo	1
1 % Valor mínimo	1
Rango 1-99%	1
Posición máxima y mínima	2
Rango completo	1
Regresión lineal	2
Error de regresión	1
Regresión cuadrática	3

Error de regresión cuadrática	1
Centroide	1
Media pico	3
Pico de distancia	2
Rango de pico	2
Rango de valle	1
Pendientes pico-valles	4
Tamaño de segmento : Media, mínimo, máximo, desviación estándar	4
Tiempo de actualización	4
Tiempo de subida	1
Tiempo de curvatura	2
Ganancia LP	1
Coeficientes LP 1-5	5
Total de funciones	55

Tabla 3. 9 Funciones estadísticas para el ComParE