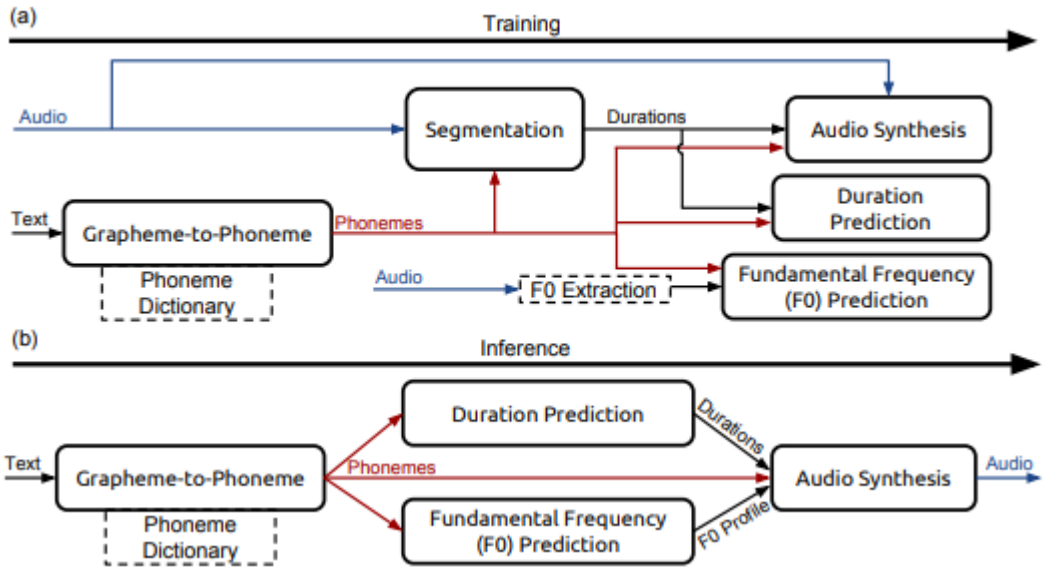
PROSODIA DE LA VOZ

Prosodia es un aspecto esencial en la síntesis de voz, ayuda a mantener expresividad e inteligibilidad de la voz. En los sistemas de síntesis nos referimos por prosodia a las componentes de duración, frecuencia fundamental, acentuación y fraseo.

En nuestro estudio este semestre nos abocamos a duración y frecuencia fundamental.

En la siguiente gráfica se muestra un ejemplo de la colocación de los elementos de las prosodia en un sintetizador (Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng & Jonathan Raiman. Deep voice: Real-time neural text-to-speech, arXiv preprint arXiv:1702.07825, 2017).



Cabe mencionar que en FastSpeech2 los audios se pueden manejar de 16 kHz a 44.1 kHz. Se convierten a 8kHz, se utilizan ventanas de 25 ms con 10 ms de offset, se obtienen ocho MFCC's por trama , con los valores delta y delta-delta, así se constituye cada vector por

tramas (Yi Ren1_, Chenxu Hu_, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech2: Fast and High-Quality End-to-End Text to Speech, 2022)

1. DURACIÓN DE LOS FONEMAS

A las repeticiones de un fonema de cada persona, se les aplica una normalización de media y variancia (CVMN, cepstral mean and variance normalization) para tener mayor robustez, finalmente se aplica una regresión lineal de máxima similitud (fMLLR, feature space Maximum Likelihood Linear Regression) para obtener valores representativos de cada fonema a partir de sus repeticiones,

Para generar la duración de fonemas, se utiliza las repeticiones de entrenamiento de cada fonema por persona y se predice la duración de ese fonema. En lugar de usar el modelo autoregresivo de FastSpeech, se utiliza la herramienta llamada alineación forzada de Montreal (Montreal Forced Aligner, MFA), con esta herramienta se logra mejorar la precisión de alineación del fonema.

El MFA es una mejora del Prosodylab-Aligner donde se destaca el uso de trifonemas. En primer término se utiliza el fonema de interés, posteriormente se usan trifonemas alrededor de ese fonema, para considerar el contexto fonético, se usan agrupamiento (clustering) para evitar la dispersión de datos. Este modelo por trifonemas es la salida de este bloque.

2. FRECUENCIA FUNDAMENTAL

La obtención de la frecuencia fundamental (F0) de un fonema ha sido un problema ya muy antiguamente abordado (60’s). La obtención en

el caso de las variaciones de la (F0) en una frase es más reciente, pero también ya había sido abordado con cierto éxito (80's).

La introducción de GPU's ha permitido renacer técnicas de alta demanda de procesamiento.

Destacan las técnicas usadas en Deep Voice en 2017 donde se predice F0 de las repeticiones en forma directa, esto ha reducido un poco la calidad de audio.

Para mejorar el contorno de FO FastSpeech2 utiliza transformada de ondeletas continua (CWT), descompone el contorno de Fo en varias escalas temporales, modela esas escalas de manera individual (una microprosodia), posteriormente las une para obtener un contorno de F0 continuo (Antti Santeri Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, and Martti Vainio, Wavelets for intonation modeling in hmm speech synthesis. In 8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013. ISCA). El resultado mejora la técnica de DeepVoice.