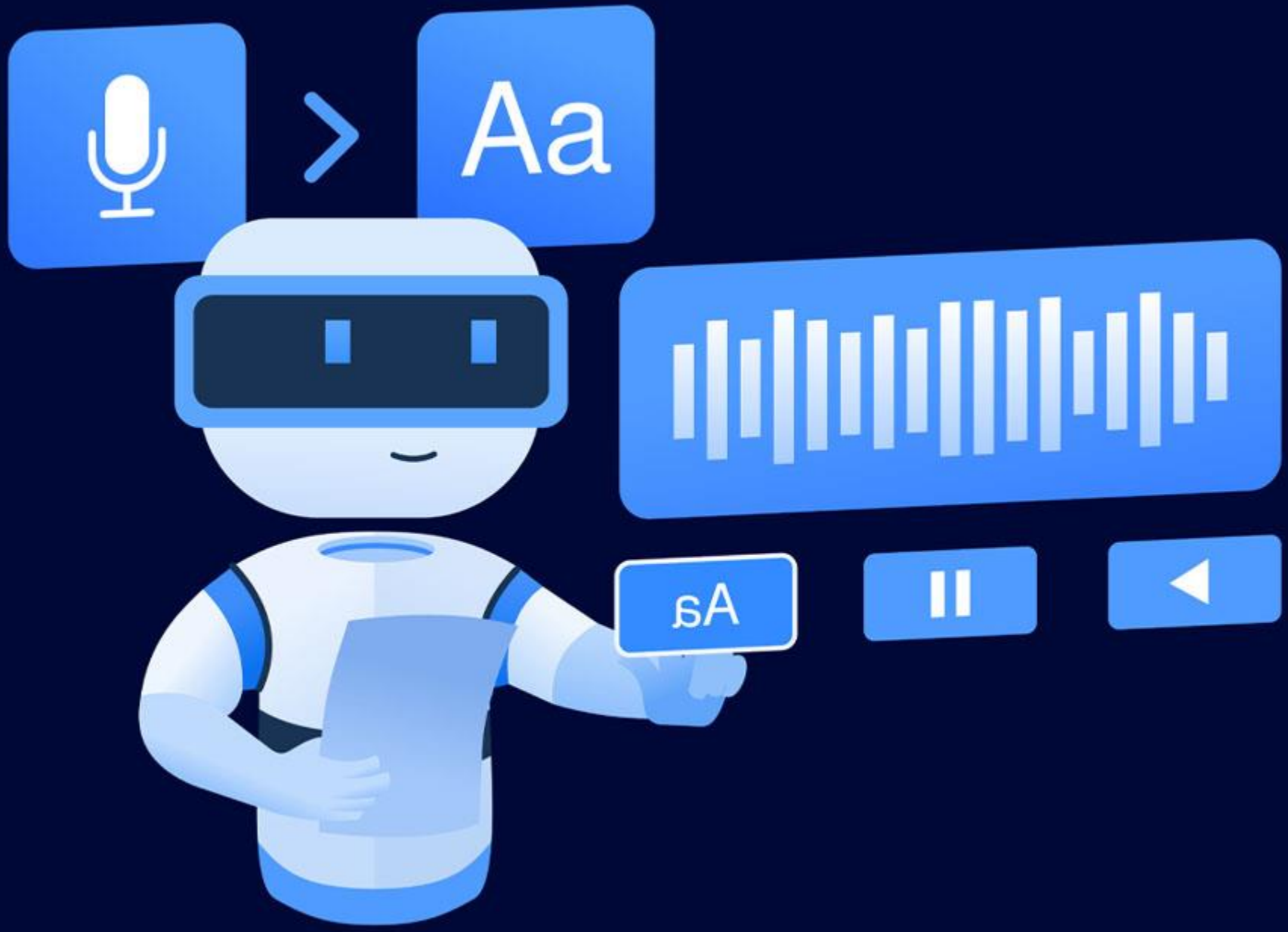
Text To speech

Franz Rivera Tellez

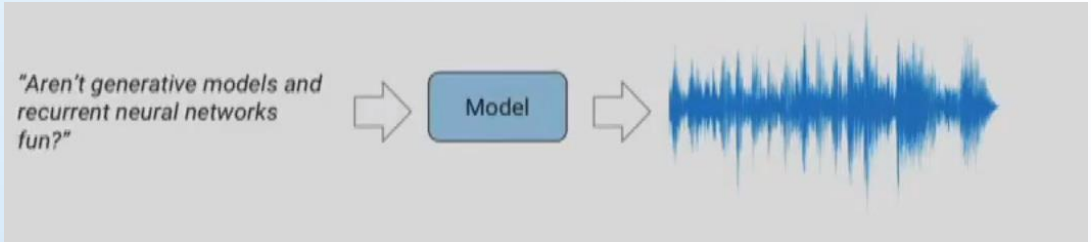

Aa
Aa

Beneficios

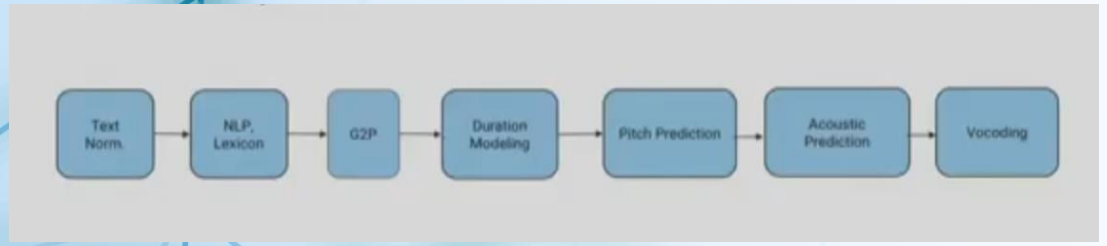
- Accesibilidad
- Experiencia de usuario
- MUltimedia


Tacotron

End to end

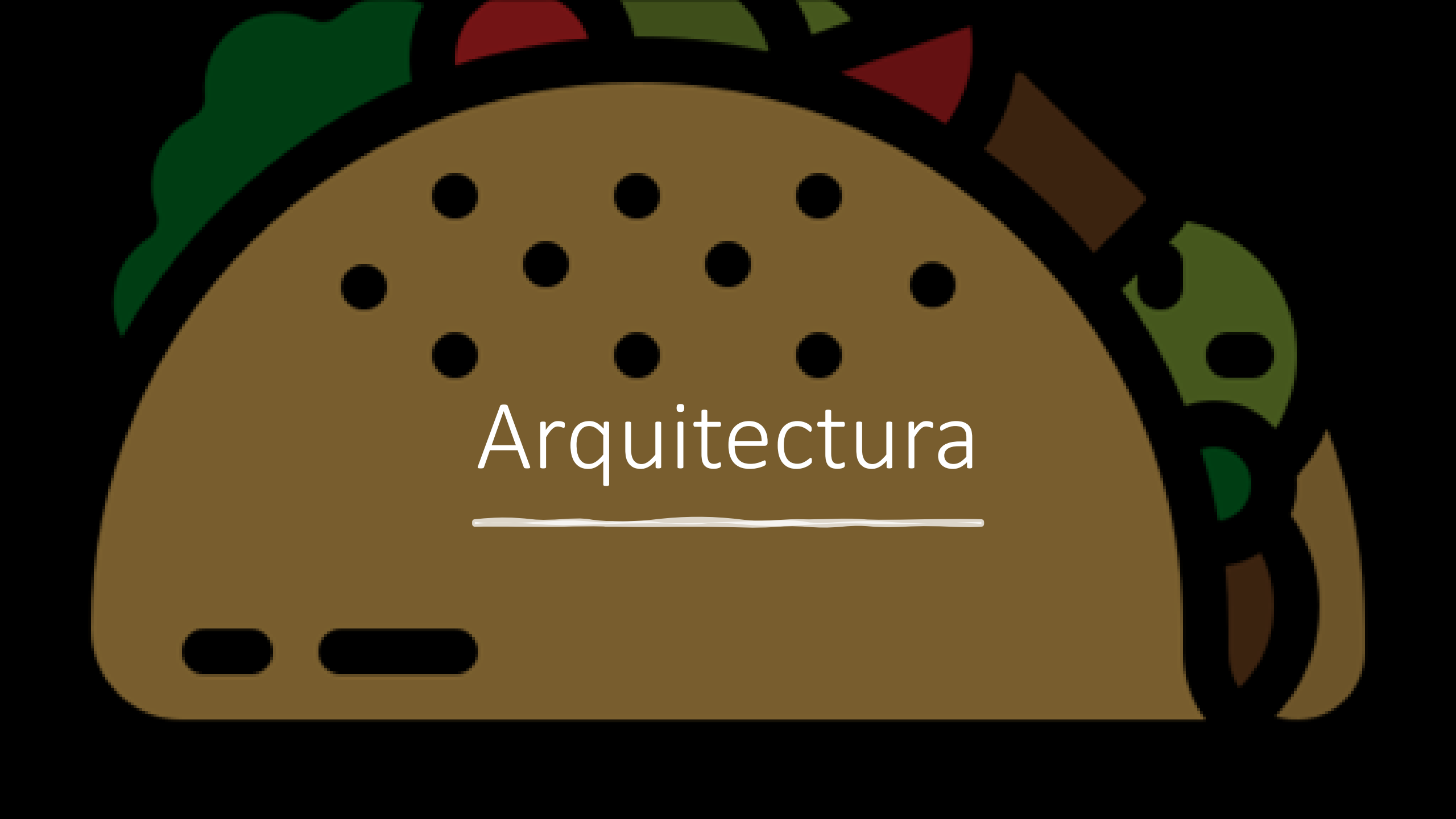


Clasic Metods



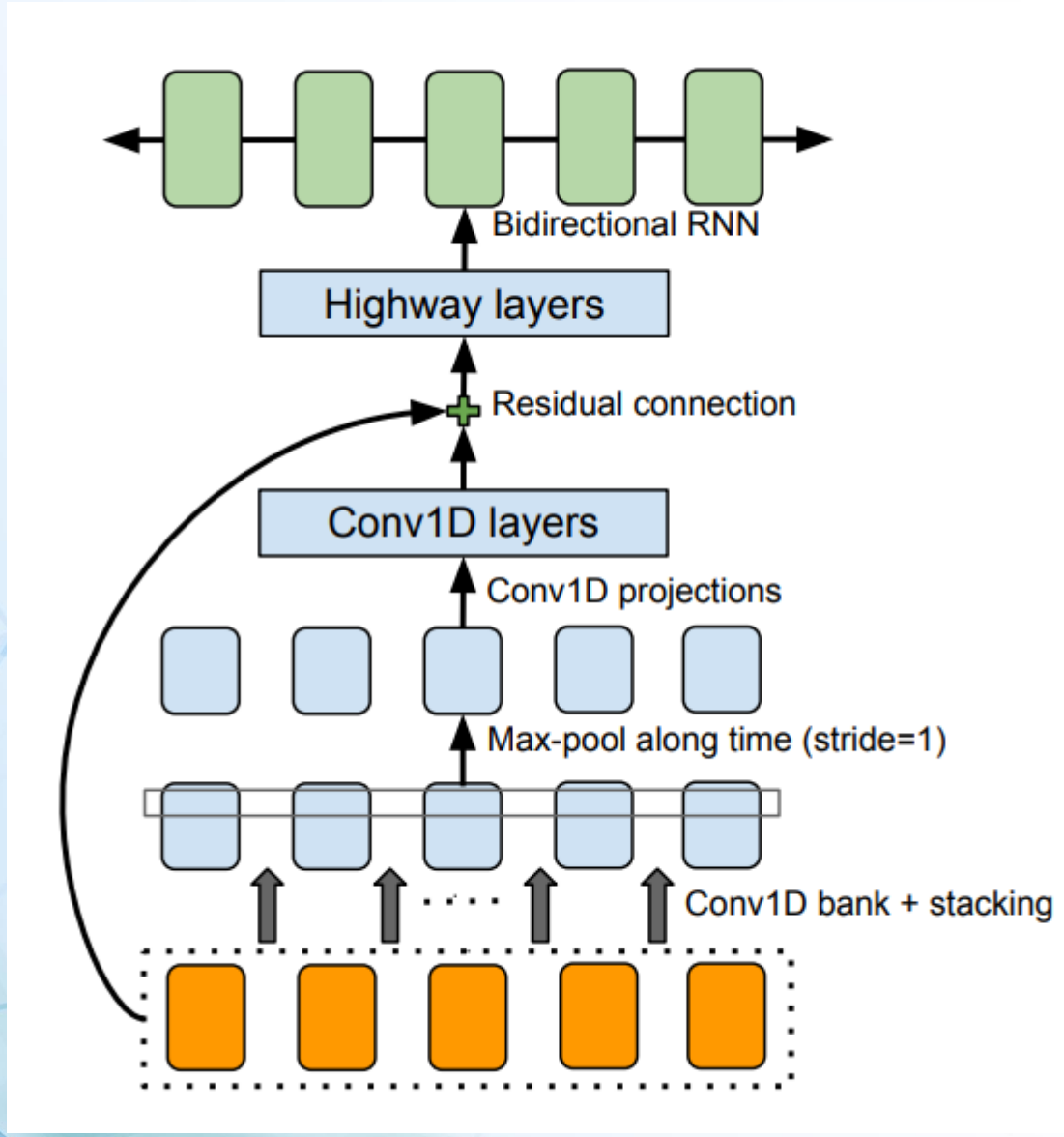
Beneficios End to end towards speech synthesis

- Simplificación del proceso
- Mayor calidad de voz
- Mayor flexibilidad y adaptabilidad
- Reducción del esfuerzo de desarrollo

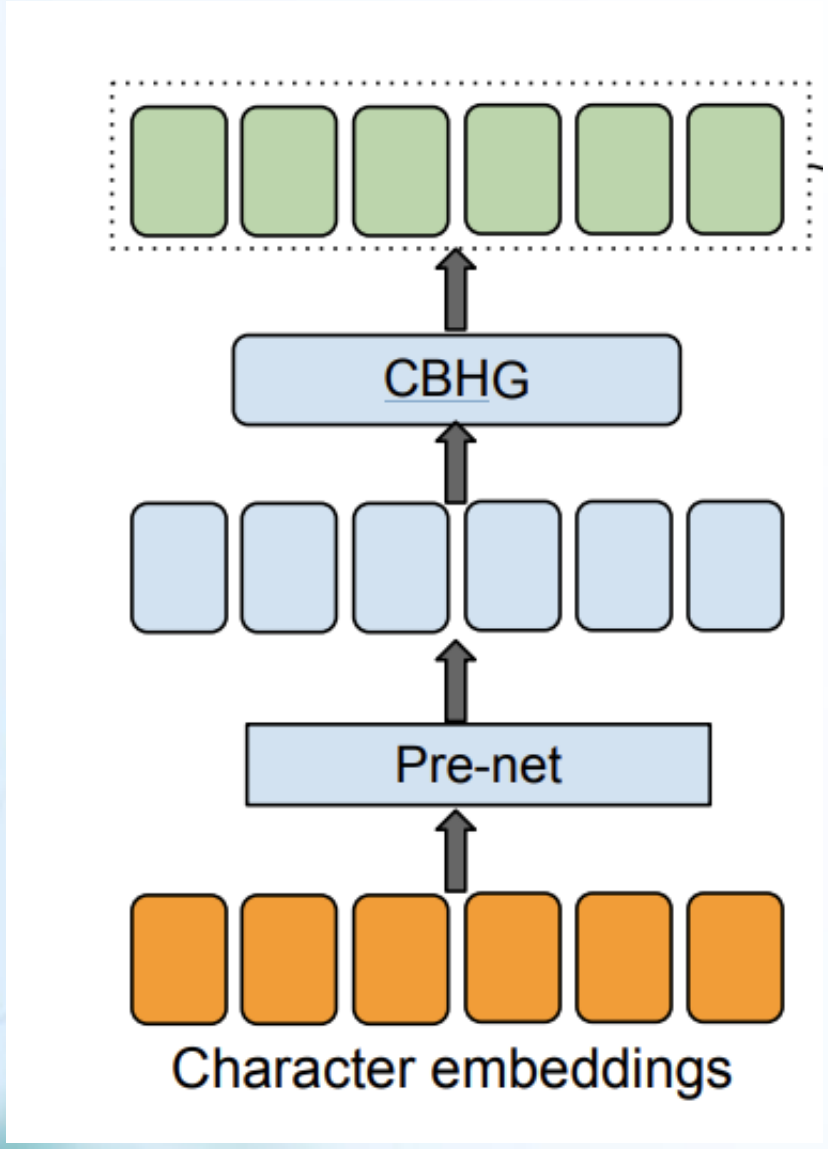


Arquitectura

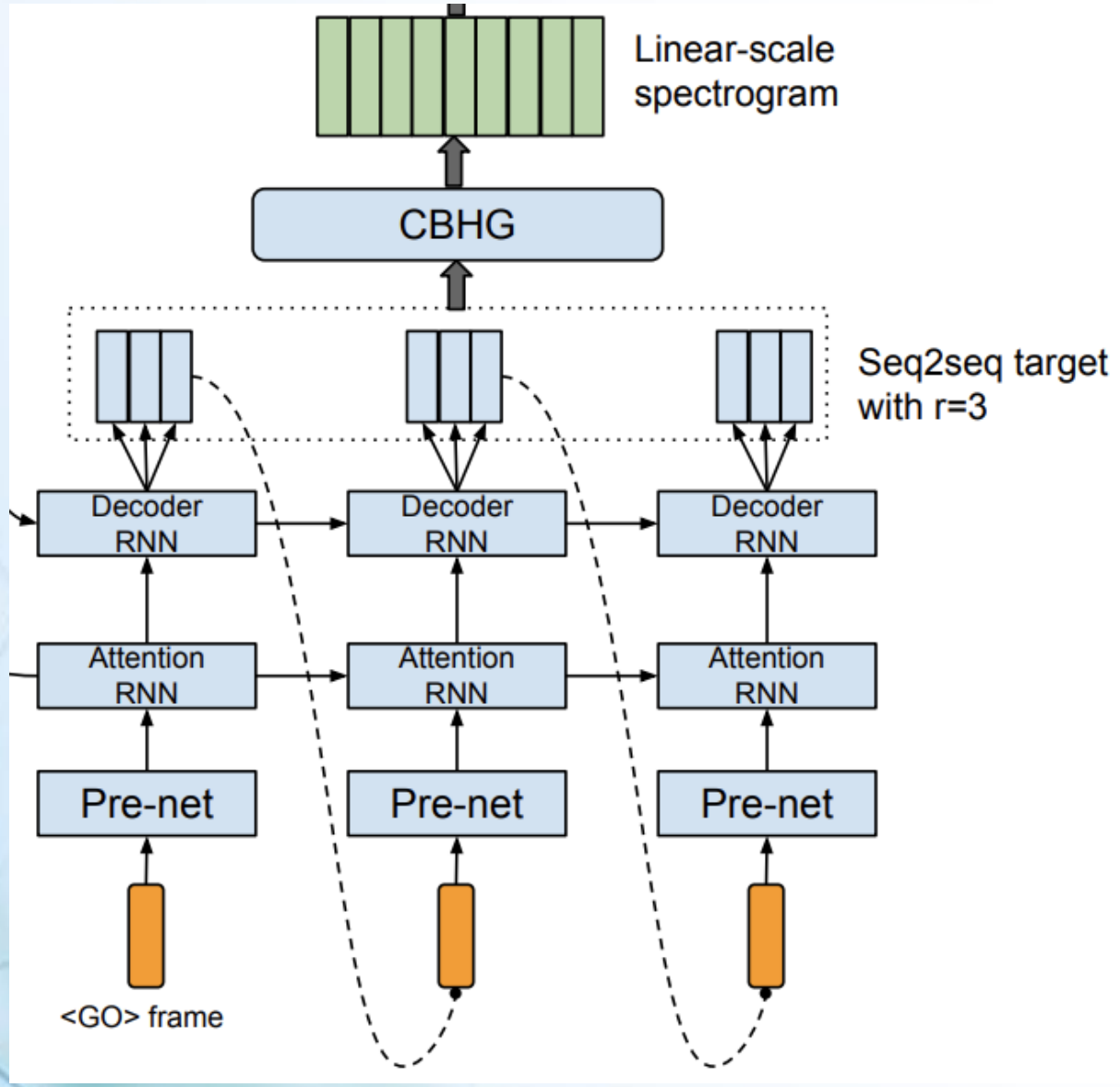
Modulo CBHG


Bidirectional RNN
Highway layers
Residual connection
Conv1D layers
Conv1D projections
Max-pool along time (stride=1)
Conv1D bank + stacking

Codificador

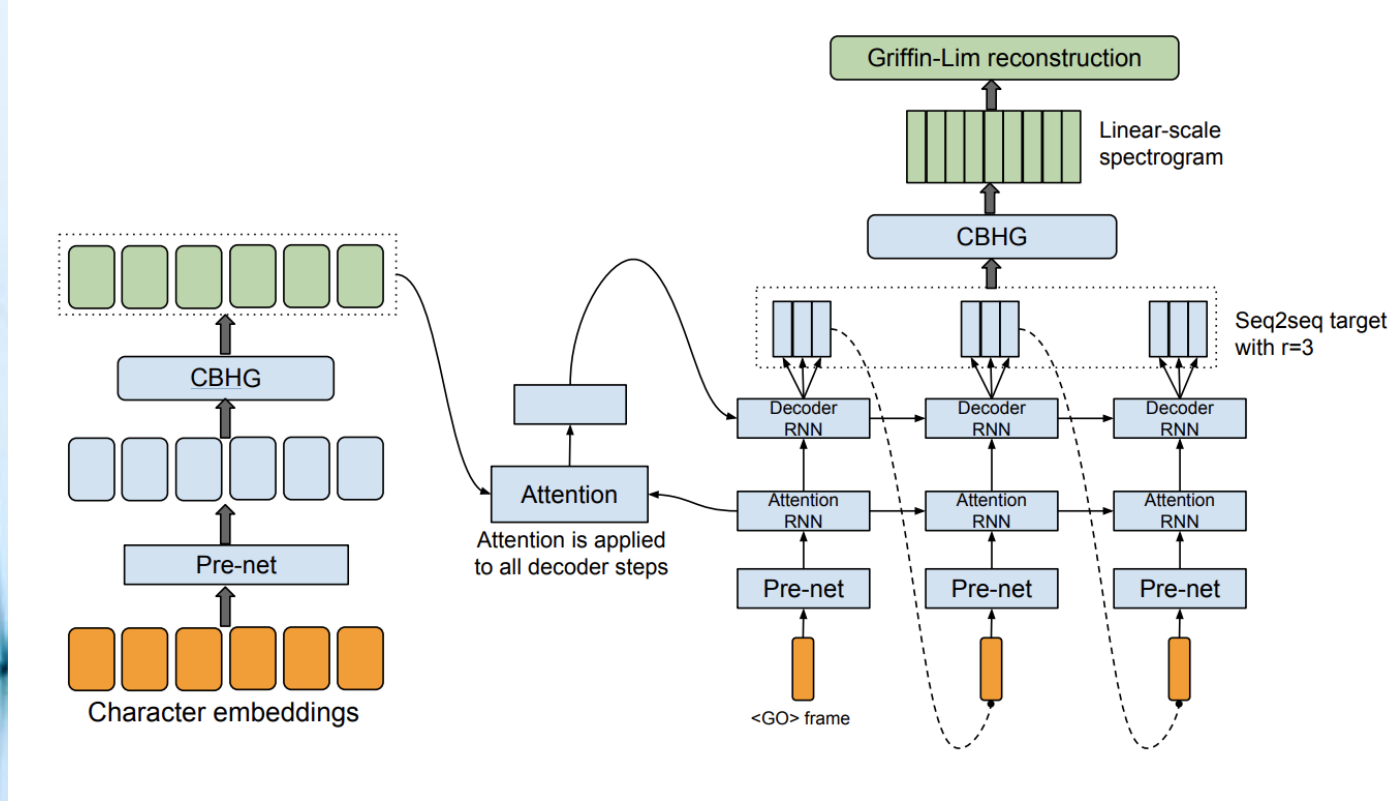


Decofificado



Post Procesamiento

Modelo completo



Entrenamiento

- Espectrograma de magnitud logarítmica con ventana Hann:
  - Se utiliza una ventana Hann para suavizar los bordes de cada trama.
  - La longitud de cada trama es de 50 ms.
  - El desplazamiento entre tramas es de 12.5 ms.
  - Se utiliza una transformada de Fourier de tamaño 2048 puntos para obtener el espectrograma.
- Preénfasis:
  - Antes de calcular el espectrograma, se aplica un filtro de preénfasis para realzar las frecuencias altas.
  - El factor de preénfasis utilizado es de 0.97.
- Tasa de muestreo:
  - La señal de audio se muestrea a una tasa de 24 kHz.

- Reducción de capa de salida:
  - Se aplica una reducción de factor "r" en la capa de salida.
  - El factor de reducción utilizado es de 2, lo que significa que la salida se reduce a la mitad de su tamaño original.
  - También se menciona que se pueden utilizar factores de reducción más grandes, como r = 5, con buenos resultados.
- Optimización:
  - Se utiliza el optimizador Adam, propuesto por Kingma y Ba en 2015, para ajustar los parámetros del modelo durante el entrenamiento.
  - La tasa de aprendizaje inicial es de 0.001.
  - La tasa de aprendizaje se reduce a 0.0005, 0.0003 y 0.0001 después de 500K, 1M y 2M pasos globales, respectivamente.

- Funciones de pérdida:
  - Se utilizan dos funciones de pérdida con pesos iguales.
  - La pérdida del decodificador Seq2seq se calcula en función del espectrograma en escala Mel.
  - La pérdida de la red de posprocesamiento se calcula en función del espectrograma en escala lineal.
- Entrenamiento:
  - El tamaño del lote de entrenamiento es de 32 muestras.
  - Todas las secuencias se rellenan con ceros hasta alcanzar una longitud máxima.
  - Aunque es común utilizar máscaras de pérdida para ignorar la pérdida en tramas rellenas con ceros, en este caso se menciona que se evitaron debido a problemas de sonidos repetidos al final.

Resultados

Table 2: 5-scale mean opinion score evaluation.

	mean opinion score
Tacotron	3.82 ± 0.085
Parametric	3.69 ± 0.109
Concatenative	4.09 ± 0.119