UNTREF UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

INGENIERÍA DE SONIDO

Desarrollo de una evaluación objetiva para sistemas de texto a voz

Tesis final presentada para obtener el título de Ingeniero de Sonido de la Universidad Nacional de Tres de Febrero (UNTREF)

TESISTA: Alejandro Sosa Welford (39912286)

TUTOR/A: Ing. Leonardo Pepino

Agosto 2022 | Caseros, Buenos Aires, Argentina

Desarrollo de una evaluación objetiva para sistemas de texto a voz

Alejandro Sosa Welford

1. Fundamentación e Introducción.

La síntesis del habla consiste en la tarea de generar una voz humana a partir de otro tipo de entrada, ya sea texto, movimiento de labios o fonemas. En la mayoría de sus aplicaciones modernas, estos sistemas toman el texto como método de entrada. Esto se debe en parte a los avances en el campo del procesamiento del lenguaje natural. Un sistema de texto a voz (TTS por sus siglas en inglés) apunta a convertir el lenguaje natural en discurso humano audible.

Históricamente esta tarea fue llevada a cabo por sistemas que concatenan fonemas pre-grabados (sistemas concatenativos) o que modelan un audio a través de parámetros acústicos definidos arbitrariamente (sistemas paramétricos). A lo largo de la última década, avances en el poder computacional permitieron explorar y desarrollar diversos modelos de TTS basados en el aprendizaje automático profundo (*Deep Learning*) a partir de diversas metodologías: En 2016 el equipo de DeepMind introduciría WaveNet [1] revolucionando el campo del TTS con el primer modelo que sintetizaba el habla humana muestra por muestra. Este pilar fue seguido por numerosos avances y mejoras basadas en sistemas de paralelización [2], transformadores [3] y sistemas de tipo *flow* [4].

Evaluar la calidad de estas distintas soluciones implica, entre otras cosas, juzgar la "naturalidad" de la voz humana generada. El estándar de oro para realizar esa evaluación son las pruebas subjetivas, realizadas sobre sistemas entrenados con bases de datos estandarizadas, usualmente en idiomas inglés o chino. El Mean Opinion Score (MOS) [5] es el método más frecuentemente utilizado para llevar a cabo esa prueba. Dicha métrica tiene un rango de 0 a 5, en la que el habla humana real yace entre las puntuaciones de 4.5 a 4.8.

Realizar un test subjetivo es costoso monetaria y temporalmente, e indefectiblemente presenta una barrera a la hora de evaluar pequeñas modificaciones o iteraciones en el desarrollo de un sistema TTS. En el presente informe se introduce un plan de investigación para llevar a cabo el desarrollo de una evaluación objetiva de la calidad de los sistemas TTS. Dicha evaluación busca tener un alto grado de correlación con los resultados de las pruebas subjetivas. Se planea ofrecer la métrica desarrollada de forma gratuita y como código abierto (*open source*).

Intercambios Transorgánicos (Dir. Gala Gonzalez Barrios) es un programa de investigaciones radicado en el Muntref Centro de Arte y Ciencia, IIAC, UNTREF. Desde este programa se realizan proyectos de investigación que desarrollan interfaces interactivas desde las artes electrónicas y las ingenierías en relación con el campo de la salud. En este momento se encuentran desarrollando un sistema TTS en español argentino, orientado a funcionar como parte de una prótesis para personas que se encuentren en la situación de comprometer su voz, parcial o totalmente. El trabajo de esta tesis busca proveer una herramienta para evaluar y ayudar al progreso de dicha investigación.

2. Objetivos General y específicos.

El objetivo general de esta investigación es desarrollar una evaluación objetiva para los sistemas de texto a voz, que tenga un alto grado de correlación con la evaluación subjetiva de tipo MOS de los mismos. Con este fín, se proponen los siguientes objetivos específicos:

- Generar y recolectar audios sintetizados por sistemas TTS de diversas calidades. Realizar distintas transformaciones (compresión, cambio de codec, agregado de reverberación y otros efectos) para variar aún más la calidad de audio recolectada.
- Diseñar y llevar a cabo una prueba subjetiva de tipo MOS donde distintos sujetos de prueba etiqueten los audios recolectados en una escala de 0 a 5, basándose en la "naturalidad" percibida de los estímulos sonoros escuchados.
- Extraer embeddings de hablante, y distintos descriptores acústicos de cada audio. Este paso implica el entrenamiento de una red neuronal para extraer los embeddings deseados. (NOTA: Aún no se han definido el tipo de embedding que serán extraídos de cada audio, posiblemente sean X-Vectors)
- Entrenar una pequeña red neuronal que aprenda a predecir el MOS de un audio a partir de su embedding de hablante correspondiente.

3. Estado del Arte.

Determinar la calidad de habla sintetizada es una problemática que atraviesa a las áreas del desarrollo de TTS, mejora del habla (*speech enhancement*), y conversión de la voz (la tarea de mantener el sentido de una oración, pero cambiar el locutor (*Voice Conversion*)). Estos sistemas pueden ser evaluados de forma tanto subjetiva como objetiva. La Distancia Cepstral de Mel (MCD por sus siglas en inglés) [6] es comúnmente utilizada para medir la calidad del habla convertida en la tarea de conversión de voz. MCD mide la distorsión de distintos rasgos acústicos de una señal, sin embargo, los mismos no siempre se ven correlacionados con la percepción humana.

Dentro del campo del *speech enhancement*, el PESQ [7] (*perceptual evaluation of speech quality* o, evaluación percibida de calidad del habla) desarrollado por ITU-T, consiste en una evaluación intrusiva para cuantificar la calidad del habla. Estima un valor MOS comparando la referencia original con la salida degradada del modelo y usando distancias paramétricas entre ambas señales. Sin embargo, en muchos casos, para sistemas de TTS no contamos con las señales originales utilizadas para entrenar una red neuronal.

En 2016, Yoshimura *et al.*[8] propuso un predictor de naturalidad en base a una red neuronal convolucional (CNN) para ser utilizada a nivel de enunciado (*utterance* o, la unidad mínima de duración de habla humana) y a nivel de sistema completo. El modelo se configuró con parámetros arbitrarios y se entrenó con una gran cantidad de resultados de evaluaciones subjetivas reales. A nivel de enunciado, esta solución presentó una gran nivel de varianza respecto a el valor MOS subjetivo real.

En 2021, Mittag y Moller [9] presentaron un evaluador de naturalidad del habla sintetizada basada en una red neuronal CNN-LSTM obteniendo resultados satisfactorios para oraciones, con pequeñas limitaciones cuando el espectro de la onda resultante se ve acotado.

4. Marco Teórico.

4.1. Técnicas posibles de alteración del hablante.

El proceso conocido como *Data augmentation* (DA), tiene el objetivo de aumentar la cantidad de información o muestras recolectadas en una base de datos, sin cambiar

fundamentalmente la etiqueta de cada muestra individual, ni tampoco modificar la distribución total de la totalidad de las muestras. Para esta investigación, poder modificar y variar incluso sutilmente las muestras recolectadas es necesario, para obtener un espacio de datos más variado. Se propone implementar técnicas de alteración del largo del tracto vocal [10] (*Voice Tract Length Perturbation,* VTLP). Dicha técnica fue desarrollada para mejorar sistemas de reconocimiento del habla exitosamente, e involucra la deformación aleatoria del espectro de cada audio. El trabajo también sugiere la posibilidad de utilizar otras técnicas de deformación temporales, y distorsiones no lineales para sintetizar una base de datos aún más robusta.

4.2. Extracción de embeddings de hablante.

La extracción de un descriptor único y útil de cada audio a evaluar es un proceso necesario para el posterior entrenamiento de la red neuronal que se encargará de calificar cada modelo TTS. En este caso los X-Vectors [11] consisten en una representación vectorizada de cada hablante que aprovecha el uso de técnicas de DA. La representación resultante ha sido útil para mejorar la eficiencia de sistemas de reconocimiento de locutores. Una implementación de la red neuronal que extrae este tipo de descriptor se encuentra disponible en el Kaldi toolkit [12].

5. Diseño de la Investigación.

El trabajo propuesto es una investigación cuantitativa de alcance exploratorio. Su propósito es el de brindar a la comunidad de investigadores que desarrollan sistemas de texto a voz, una evaluación objetiva automatizada que presente un alto grado de correlación con las pruebas subjetivas que conforman el estándar de la industria para juzgar el habla. Se plantea extraer un descriptor de cada audio a juzgar, y entrenar una pequeña red neuronal de forma supervisada, de modo que la misma pueda predecir el valor MOS que obtendría el audio si fuese juzgado subjetivamente por un grupo de individuos.

Para el desarrollo de la investigación se plantean las siguientes etapas de trabajo. Las mismas están sujetas a revisiones y modificaciones en base a la evolución del proyecto.

- 1° etapa: **Recolección de audios sintetizados.**

  La primera instancia de la investigación consiste en recolectar audios sintetizados por sistemas TTS de variada calidad de clonado de voz y procedencia, además de audios de hablantes humanos reales. Se decide que la investigación se conducirá en el idioma español.

- 2° etapa: **Transformación de audios recolectados.**

  Con el objetivo de variar la calidad de los audios recolectados, herramientas para alterar los mismo como Vocal Tract Length Perturbation serán utilizadas. También los audios pueden ser comprimidos, atravesar distintos codecs o ser reverberados con este mismo objetivo.

- 3° etapa: **Extracción de descriptores objetivos de cada audio.**

  Se propone extraer X-VECTORS de cada audio. Estos embeddings son utilizados para reconocimiento de hablantes. Estos descriptores se extraen mediante una red neuronal que deberá ser configurada y posiblemente re-entrenada para funcionar

con el idioma español.

• 4° etapa: **Diseño de una prueba subjetiva para etiquetar los audios.**

Con el propósito de obtener una puntuación MOS para cada audio recolectado, se llevarán a cabo pruebas subjetivas para obtener estas etiquetas. Seguido a esto, una validación de los datos obtenidos también deberá ser realizada.

• 5° etapa: **Diseño de una red neuronal para predecir la naturalidad de cada audio.**

Se entrenará una pequeña red neuronal de forma supervisada, con los audios recolectados como entrada y sus calificaciones MOS como salida deseada. La misma contará probablemente con 2 a 3 capas completamente conectadas, seguidas de una activación que aún no ha sido determinada, que indicará la predicción de cada inferencia. La función de costo y el ajuste de la red tendrán como objetivo acercar sus predicciones a los valores correctos MOS recolectados. Para poner a prueba el modelo entrenado, se reserva una parte del conjunto de datos recolectados para llevar a cabo una evaluación.

5.1. Diseño prueba subjetiva: Encuesta y Muestra.

El diseño de la prueba subjetiva se basa en las especificaciones provistas por las recomendaciones del estándar ITU-T Rec. P.807. Todos los sujetos encuestados deberán cumplir con la condición de ser normo-oyentes. La encuesta podría ser conducida en la página de neuropruebas.org desarrollada por el Laboratorio de Inteligencia Artificial Aplicada, ICC - UBA - CONICET. Esta página permite que usuarios completen test subjetivos de forma asincrónica. Es ideal para evaluaciones que requieren de mínimo entrenamiento como la que se propone para la investigación en cuestión. El test consistirá de una serie de audios que deberán ser evaluados en una escala de tipo Likert de 5 puntos por cada sujeto. La cantidad de audios a evaluar no estará acotada en principio, permitiendo a cada usuario estudiar cuantos audios quiera. El propósito de la encuesta subjetiva es la de etiquetar los audios recolectados previamente, con una puntuación. La cantidad de etiquetas necesarias están determinadas por el entrenamiento de la red neuronal que se desarrollará a posteriori. Un precedente útil puede ser tomado del trabajo de Deja *et al.*[13] en cual se llevó a cabo una metodología similar. Sujeto a la cantidad de audios que evalúe cada persona, se propone que en principio serán necesarios alrededor de 100 sujetos de prueba. Este número es aproximado y será revisado a la hora de plantear la arquitectura de la red neuronal que evaluará objetivamente cada audio a partir de su representación vectorizada.

6. Análisis de los resultados: Aplicaciones estadísticas.

Una vez entrenada la red neuronal, se deberá evaluar su rendimiento con una serie de datos a los que no haya sido expuesta previamente. Dichas inferencias se corresponden con una serie de resultados en una escala de 0 a 5. Los mismos deberán ser correlacionados con los resultados “correctos”, obtenidos a partir de las evaluaciones subjetivas realizadas previamente. Esta métrica se calculará a partir del coeficiente de correlación de Pearson.

7. Conclusiones.

La investigación planeada involucraría la recolección de audios de distintos sistemas TTS y expansión de esa base de datos por medio de técnicas de DA. Luego, el diseño e implementación de un test subjetivo de tipo MOS para etiquetar esos audios recolectados con una puntuación, para utilizarse en el desarrollo y entrenamiento de una red neuronal que pueda evaluar la calidad de un audio. Finalmente se llevaría a cabo la correlación de los resultados de distintas inferencias realizadas por la red entrenada, con puntuaciones MOS subjetivas para evaluar el grado de precisión del modelo desarrollado.

8. Líneas futuras de investigación.

Al finalizar el desarrollo, de acuerdo a los resultados alcanzados, se podría evaluar el modelo realizado en otros idiomas para verificar su funcionamiento multilingüe. Otro posible desarrollo será obtener una base de datos más robusta, con un número de encuestados mayor, para re-entrenar y refinar el funcionamiento de la red neuronal. También se plantea la posibilidad de empaquetar el modelo para poder ser utilizado como una librería de Python disponible como código abierto, para facilitar su uso en producción de sistemas TTS.

9. Cronograma.

El siguiente diagrama de Gantt enuncia el cronograma de actividades y tareas previstas al momento de la realización de la investigación.

Etapa	Actividad	Quincenas															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
E1	Definición del proyecto y redacción del plan de investigación	X															
E1	Revisión de la literatura	X	X	X	X	X	X	X									
E2	Recolección de audios de sistemas TTS		X	X	X												
E2	Implementación de técnicas de alteración de hablante		X	X	X												
E3	Diseño de la prueba subjetiva			X	X	X											
E3	Evaluación subjetiva						X	X	X	X							
E4	Extracción de embeddings de hablante y diseño de red neuronal						X	X	X	X							
E4	Entrenamiento y refinación del modelo								X	X	X	X					
E5	Obtención, análisis y conclusión de resultados												X	X			
E6	Redacción de tesis													X	X	X	X

10.Bibliografía.

[1] Oord, Aaron van den and Dieleman, Sander and Zen, Heiga and Simonyan, Karen and Vinyals, Oriol and Graves, Alex and Kalchbrenner, Nal and Senior, Andrew and Kavukcuoglu, Koray. WaveNet: A Generative Model for Raw Audio, arXiv (2016).
[2] Oord, Aaron van den and Li, Yazhe and Babuschkin, Igor and Simonyan, Karen and Vinyals, Oriol and Kavukcuoglu, Koray and Driessche, George van den and Lockhart, Edward and Cobo, Luis C. and Stimberg, Florian and Casagrande, Norman and Grewe, Dominik and Noury, Seb and Dieleman, Sander and Elsen, Erich and Kalchbrenner, Nal and Zen, Heiga and Graves, Alex and King, Helen and Walters, Tom and Belov, Dan and Hassabis, Demis. Parallel WaveNet: Fast High-Fidelity Speech Synthesis, arXiv (2017).
[3] Ren, Yi and Ruan, Yangjun and Tan, Xu and Qin, Tao and Zhao, Sheng and Zhao, Zhou and Liu, Tie-Yan. FastSpeech: Fast, Robust and Controllable Text to Speech, arXiv (2019).
[4] Prenger, Ryan and Valle, Rafael and Catanzaro, Bryan. WaveGlow: A Flow-based Generative Network for Speech Synthesis, arXiv (2018).
[5] ITU-T Rec. P.800. Methods for subjective determination of transmission quality (1996).
[6] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment, Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing (1993).
[7] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ) a new method for speech quality assessment of telephone networks and codecs, Proc. ICASSP (2001)
[8] T. Yoshimura, G. Eje Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda. A hierarchical predictor of synthetic speech naturalness using neural networks, Proc. Interspeech (2016).
[9] Gabriel Mittag and Sebastian Möller. Deep Learning Based Assessment of Synthetic Speech Naturalness, Interspeech 2020 ISCA (2020).
[10] Navdeep Jaitly and E. Hinton. Vocal Tract Length Perturbation (VTLP) improves speech recognition, Proc.of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing, (2013).
[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur. X-Vectors: Robust DNN Embeddings for Speaker Recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018).
[12] Daniel Povey. Kaldi Speech Recognition Toolkit. https://github.com/kaldi-asr/kaldi.
[13] Kamil, Deja and Ariadna, Sanchez and Julian, Roth and Marius, Cotescu. Automatic Evaluation of Speaker Similarity, arXiv (2022).