Generación y Modificación de Archivos .ges para VocalTractLab

1. Introducción

El presente informe documenta el desarrollo y análisis de un conjunto de funciones en Python destinadas a la generación y modificación de archivos .ges, utilizados por la herramienta de simulación articulatoria VocalTractLab (VTL) versión más actualizada para Windows. Esta herramienta permite modelar la producción del habla a partir de parámetros articulatorios y acústicos definidos en estructuras XML.

Este módulo fue desarrollado con el objetivo de facilitar la creación automatizada de archivos .ges para diferentes vocales, así como permitir modificaciones articulatorias específicas que afectan directamente la simulación del tracto vocal humano.

2. Fundamento Teórico

La representación de gestos articulatorios no se define directamente mediante coordenadas o valores numéricos en el archivo **.ges**. En cambio, se hace referencia a etiquetas predefinidas en el archivo **.speaker**, el cual contiene las formas articulatorias estandarizadas para vocales y otras configuraciones del tracto vocal.

Esto implica que cualquier intento de modificar valores numéricos de forma directa debe hacerse con extrema precaución, ya que podría violar la lógica interna del motor de simulación de VTL y provocar errores en la generación de audio o animaciones articulatorias.

VocalTractLab (VTL) es un simulador de producción de habla que permite visualizar y sintetizar el comportamiento dinámico del tracto vocal y la glotis. La arquitectura de VTL se apoya fundamentalmente en dos archivos clave: el archivo de configuración del hablante (**.speaker**) y la puntuación gestual (**.ges**).

- **Archivo .speaker: Modelo Anatómico del Hablante**

  El archivo **.speake**r es un archivo en formato XML que contiene:

  - **Modelo del tracto vocal supraglotal**: estructuras anatómicas, parámetros deformables y una lista de **formas (shapes) etiquetadas**, que representan configuraciones articulatorias como vocales y consonantes.

  - **Modelos glotales (glottis_models)**: definen distintos patrones de fonación, mediante parámetros estáticos y de control.

  Estas formas predefinidas son utilizadas como **objetivos articulatorios**. Las etiquetas como "a", "i", "ll-labial-closure" o "modal" son referencias clave para los gestos que se especifican en el .ges.

- **Archivo .ges: Puntuación Gestual**

  El archivo .ges es también un XML que describe la **secuencia temporal de gestos articulatorios** que forman una elocución. Se organiza en ocho capas (tiers), que incluyen:

  - Vocales, labios, punta y cuerpo de la lengua
  - Gestos velofaríngeos
  - Forma glotal
  - F0
  - Presión pulmonar

  Cada gesto en estas capas no define directamente los parámetros articulatorios, sino que **apunta mediante su etiqueta** a una de las formas definidas en el archivo **.speaker**. Por ejemplo, un gesto para la vocal "i" apuntará a la forma "i" definida en el **.speaker**.

- **Interacción entre .speaker y .ges**

  La conexión entre estos archivos es **referencial**:

  - Un gesto supraglótico en **.ges** debe indicar una **etiqueta de forma**, no parámetros numéricos.
  - Al ejecutar una simulación, VTL interpreta el gesto como una transición hacia la configuración almacenada en el **.speaker**, usando modelos dinámicos basados en duración y constante de tiempo.

3. Objetivos del Código

- Generar archivos **.ges** válidos para diferentes vocales.
- Asignar valores controlados a parámetros acústicos como la frecuencia fundamental (F0) y la presión pulmonar.
- Incorporar, opcionalmente, gestos adicionales como la apertura de la mandíbula, el ensanchamiento labial y la posición de la punta de la lengua.
- Adaptar el código al modelo de control gestual de VTL, cuidando la referencia a etiquetas definidas en **.speaker**.

4. Descripción del Código

a. Estructura General

El módulo está compuesto por varias funciones clave:

- create_vowel_ges(...): genera un archivo .ges que contiene gestos para una vocal específica, así como gestos glotales, de F0 y presión pulmonar.

- generate_multiple_vowel_ges(...): automatiza la creación de archivos para todas las vocales disponibles (["a", "e", "i", "o", "u"]).

- generate_single_vowel_ges(...): permite la generación individual para una vocal especificada.

- modify_vocal_tract_position(...): modifica un archivo .ges existente para introducir gestos adicionales relacionados con la configuración física del tracto vocal.

b. Validación de Etiquetas

Antes de la generación, se valida que la vocal solicitada esté incluida en la lista **valid_shapes**, que representa las etiquetas aceptadas por el archivo **.speaker.** Esta validación es fundamental para asegurar la compatibilidad con el modelo acústico-articulatorio de VTL.

c. Estructura XML

Los gestos se representan mediante estructuras XML del tipo:

<gesture_sequence type="vowel-gestures">

<gesture value="a" slope="0.000000" duration_s="0.600000" time_constant_s="0.020000" neutral="0"/>

</gesture_sequence>

Se incluyen cuatro gestos principales por defecto:

- vocales (vowel-gestures)

- forma glotal (glottal-shape-gestures)

- frecuencia fundamental (f0-gestures)

- presión pulmonar (lung-pressure-gestures)

Adicionalmente, mediante modify_vocal_tract_position, se pueden agregar gestos articulatorios como:

- apertura mandibular (jaw-gestures)
- ancho labial (lip-gestures)
- posición de la lengua (tongue-tip-gestures)

d. Escritura del Archivo

Los archivos se guardan en formato .ges utilizando xml.dom.minidom para garantizar una escritura legible (indentada) del XML.

```
with open(output_filename, 'w', encoding='utf-8') as f:
    f.write(pretty_xml)
```

5. Consideraciones Técnicas y Precauciones:
   a. Compatibilidad con VTL

      Dado que el motor de simulación de VTL espera etiquetas articulatorias en lugar de valores numéricos para ciertos gestos (por ejemplo, vocales y formas glotales), cualquier modificación debe realizarse dentro del marco permitido por el archivo **.speaker**. Usar valores no definidos puede resultar en errores de simulación o archivos inválidos.

   b. Riesgo de Modificaciones Directas

      La función **modify_vocal_tract_position** agrega gestos adicionales basados en valores numéricos. Esta funcionalidad es útil para experimentos, pero se debe tener precaución al usarla, ya que no todos los valores numéricos son compatibles con las restricciones biomecánicas del modelo articulatorio subyacente en VTL.

6. **Observaciones:**

Cualquier intento de modificar directamente parámetros como jaw_height, lip_width, etc., dentro del archivo .ges, **viola la estructura interna de VTL**. Para realizar cambios válidos, deben seguirse dos caminos:

1. **Modificar el archivo .speaker**, creando nuevas formas con las configuraciones deseadas.
2. **Exportar y manipular archivos derivados**, como .tract o .tube, que representan estados intermedios del tracto vocal.

Procesamiento y Conversión de Archivos .ges a Audio y Datos

El módulo presentado interactúa con **VocalTractLab (VTL)**, permitiendo la simulación articulatoria del habla a partir de gestos definidos en el archivo .ges y referenciados en el archivo .speaker.

Durante la conversión de un gesto articulatorio en señal de audio:

- Se utiliza la función vtlGesturalScoreToAudio() de VTL para transformar los gestos en una señal acústica (.wav).
- Se generan archivos auxiliares (.misc y .csv) con información estructurada sobre los gestos y parámetros.

Objetivos del Código

- **Convertir archivos .ges en audio (.wav)** utilizando la API de VocalTractLab.
- **Generar archivos .misc y .csv** para el análisis detallado de los parámetros gestuales utilizados en la simulación.
- **Automatizar la conversión de múltiples archivos .ges** dentro de una carpeta, permitiendo el procesamiento en lote.

Descripción del Código

a. **Estructura General**

   El módulo está compuesto por varias funciones clave:

   - generate_wav_and_csv_from_ges(...): toma un archivo .ges y genera .wav, .misc y .csv.
   - generate_wavs_and_csvs_from_multiple_ges(...): procesa en lote todos los archivos .ges en una carpeta.
   - convert_misc_to_csv(...): transforma un archivo .misc en .csv.

b. **Interacción con VocalTractLab**

   Antes de realizar la conversión, el código valida que la API de VTL pueda cargarse correctamente. Luego, inicializa la simulación con el archivo .speaker, asegurando que los gestos del archivo .ges correspondan a etiquetas reconocidas por VTL.

c. **Generación de Archivos**

   El proceso de conversión involucra:

   1. **Inicialización de la API VTL** con el modelo articulatorio del hablante.
   2. **Procesamiento de gestos** con vtlGesturalScoreToAudio(), transformándolos en audio (.wav).
   3. **Extracción de datos estructurados**, guardando la información en .misc y luego en .csv mediante numpy y pandas.

d. **Escritura del Archivo CSV**

   Los datos del archivo .misc se convierten a .csv con una estructura tabular.

Consideraciones y Precauciones

- **Compatibilidad con VTL**

  Dado que la simulación en VocalTractLab se basa en referencias a etiquetas en el archivo .speaker, cualquier conversión debe seguir la lógica interna del motor de VTL. Modificaciones arbitrarias pueden generar errores en la síntesis de audio.

Bibliografías utilizadas:

- Birkholz, P. (2013). *Modeling consonant-vowel coarticulation for articulatory speech synthesi*. PLoS ONE.
- *GitHub - TUD-STKS/VocalTractLab-dev: VocalTractLab development repo*. (s/f). GitHub. Recuperado el 20 de mayo de 2025, de https://github-com.translate.goog/TUD-STKS/VocalTractLab-dev?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es-419&_x_tr_pto=sc
- *VocalTractLab*. (s/f). Vocaltractlab.de. Recuperado el 20 de mayo de 2025, de https://www.vocaltractlab.de/
- *VocalTractLab-Python: Articulatory (text-to-) speech synthesis for Python*. (s/f).
- (S/f). Github.com. Recuperado el 20 de mayo de 2025, de https://github.com/quantling/create_vtl_corpus/blob/main/create_vtl_corpus/generate_data.py
- Fowler, A. (2015). *NoSQL For Dummies* (1a ed.). John Wiley & Sons.