# Ingeniería de la Naturalidad: Protocolos Avanzados de Maquetación y Preprocesamiento de Manuscritos para la Síntesis de Voz Neural de Alta Fidelidad

## 1. Introducción: La Convergencia de la Lingüística Computacional y la Producción Editorial

La industria de la publicación de audiolibros ha experimentado una transformación tectónica en la última década, impulsada fundamentalmente por la evolución de los motores de síntesis de voz (TTS, por sus siglas en inglés) desde modelos concatenativos rígidos hacia arquitecturas de redes neuronales profundas (DNN). Históricamente, la producción de un audiolibro era un proceso lineal y analógico: un manuscrito se entregaba a un narrador humano, quien interpretaba las señales visuales (puntuación, formato, contexto semántico) y las traducía en señales acústicas. En este paradigma tradicional, el "formato" del documento de Word era irrelevante siempre que fuera legible para el ojo humano. Un narrador podía ignorar un doble espacio accidental, corregir mentalmente una errata obvia o interpretar una tabla compleja sin detener la grabación.

Sin embargo, la adopción masiva de voces neurales, como Amazon Polly (NTTS), Azure Neural Voice y ElevenLabs, ha invertido esta relación.1 Aunque estos sistemas son capaces de generar una prosodia asombrosamente humana —capturando matices de entonación, ritmo y timbre que antes eran imposibles— carecen de la capacidad de improvisación y corrección de errores de un ser humano. Para un motor neural, el documento de entrada no es solo una historia; es una secuencia estricta de instrucciones de control. La calidad del audio resultante, por tanto, ya no depende únicamente de la sofisticación del modelo de IA, sino de la higiene, estructura y semántica del archivo de origen.

El principio fundamental que rige la producción de audiolibros con IA es que la "naturalidad" no es un accidente, sino el resultado de una ingeniería documental precisa. Un archivo de Microsoft Word (.docx) mal maquetado —lleno de caracteres ocultos, estilos inconsistentes o ambigüedades tipográficas— introduce ruido en el modelo de predicción de la red neuronal, resultando en artefactos auditivos conocidos como el "valle inquietante" sonoro: pausas antinaturales, entonaciones erráticas y una falta de coherencia respiratoria.4

Este informe técnico proporciona una guía exhaustiva para autores, editores e ingenieros de audio sobre cómo estructurar, limpiar y enriquecer archivos de Word para la conversión a audiolibros neurales. Abordaremos desde la arquitectura XML subyacente de los archivos .docx hasta la inyección estratégica de SSML (Speech Synthesis Markup Language) a través de estilos de párrafo, garantizando que el producto final cumpla con los rigurosos estándares técnicos de plataformas como ACX/Audible y ofrezca una experiencia auditiva indistinguible de una narración humana profesional.7

### 1.1 La Arquitectura Cognitiva de la Voz Neural

Para comprender cómo formatear un documento, primero debemos entender cómo "lee" la máquina. A diferencia de los sistemas TTS antiguos que ensamblaban fonemas pregrabados (síntesis concatenativa), los motores neurales modernos utilizan modelos de secuencia a secuencia (Seq2Seq) que generan espectrogramas de mel completos a partir del texto de entrada.1

Esto tiene una implicación crítica para la maquetación: el **Contexto**. El motor neural analiza ventanas de texto (oraciones o párrafos completos) para determinar la entonación adecuada. Si el formato del documento rompe artificialmente este contexto —por ejemplo, mediante un salto de línea manual en medio de una frase— la red neuronal interpretará el final de la línea como el final de una unidad semántica, provocando una caída de tono (pitch drop) característica de un final de oración, incluso si la frase gramatical continúa. Por lo tanto, la integridad estructural del párrafo en Word se convierte en el factor determinante de la fluidez prosódica.8

## 2. Anatomía Técnica del Archivo Fuente (.docx) y Limpieza de Datos

El formato .docx no es un archivo plano; es un contenedor comprimido (ZIP) que alberga una colección de archivos XML. El contenido real reside en document.xml, mientras que la apariencia visual se define en styles.xml.10 Para la síntesis de voz, la apariencia visual es irrelevante; lo que importa es la estructura semántica. Muchos "errores" de audio provienen de confundir la presentación visual con la estructura lógica.

### 2.1 Higiene Tipográfica y Eliminación de Artefactos (Pre-procesamiento)

El primer paso en cualquier flujo de trabajo de producción de audiolibros neurales es la "sanitización" del manuscrito. Los procesadores de texto modernos están diseñados para la impresión en papel, introduciendo caracteres que son estéticamente agradables pero acústicamente destructivos.

#### 2.1.1 La Plaga de los Espacios en Blanco (Whitespace Management)

En la tipografía digital para impresión, el espaciado es flexible. En TTS, el espaciado es tiempo.

* **Dobles Espacios:** Una práctica heredada de las máquinas de escribir es colocar dos espacios después de un punto. Visualmente es inocuo. Acústicamente, algunos motores TTS interpretan esto como una instrucción para extender la pausa de silencio entre oraciones más allá del estándar de 500ms, rompiendo el ritmo natural.11
* **Espacios de No Separación (Non-breaking spaces):** A menudo se usan para mantener juntas cifras y unidades (ej. "10 kg"). Si el motor TTS no está configurado para normalizar esto, puede ignorar el espacio y leer "diezquilogramos" apresuradamente, o peor, leer el código del carácter si la codificación falla.
* **Tabulaciones:** Las tabulaciones (^t en la búsqueda de Word) se usan para alinear texto visualmente. Para el TTS, son caracteres nulos o pausas impredecibles.

Protocolo de Limpieza:

Se debe ejecutar una rutina de limpieza estricta utilizando "Buscar y Reemplazar" en Word o expresiones regulares (Regex) en un script de pre-procesamiento 13:

1. Reemplazar ^w^w (doble espacio) por ^w (espacio simple) repetidamente hasta que el contador sea cero.
2. Eliminar todas las tabulaciones (^t).
3. Eliminar espacios al inicio y final de los párrafos, ya que pueden confundir a los algoritmos de detección de inicio de voz.

#### 2.1.2 Saltos de Línea Manuales vs. Saltos de Párrafo

Este es el error más crítico en la maquetación amateur.

* **Salto de Párrafo (Hard Return, Enter, ^p):** Indica el fin de una idea y el comienzo de otra. Los motores neurales, como Amazon Polly, insertan automáticamente una pausa significativa (aprox. 500ms - 1s) y resetean el contorno de entonación.15
* **Salto de Línea Manual (Soft Return, Shift+Enter, ^l):** Fuerza al texto a bajar de línea sin cerrar el párrafo lógico. Visualmente parece un párrafo nuevo o un ajuste de línea poético. Sin embargo, acústicamente, el motor puede tratarlo como un espacio simple o, peor, como una ruptura abrupta sin la pausa adecuada, colapsando la prosodia.

**Análisis de Impacto:** Si un autor usa saltos manuales para formatear diálogos, el motor leerá el diálogo como una sola frase atropellada o con pausas erráticas. Es imperativo convertir todos los saltos manuales (^l) en saltos de párrafo reales (^p) a menos que sea poesía intencional donde el encabalgamiento (enjambment) sea deseado.16

#### 2.1.3 Caracteres Invisibles y Metadatos

Los archivos Word a menudo acumulan "basura" invisible al copiar y pegar desde la web o PDFs, como marcadores de dirección de texto (LTR/RTL marks) o guiones discrecionales. Estos caracteres pueden causar que el motor TTS se detenga, tartamudee o falle completamente en la síntesis. Herramientas de limpieza de texto basadas en Regex son esenciales para purgar cualquier carácter fuera del rango ASCII estándar o del set extendido necesario para el idioma específico.17

| **Carácter Visual** | **Representación Word** | **Interpretación TTS Neural** | **Acción Recomendada** |
| --- | --- | --- | --- |
| Espacio Doble | .. (dos puntos visuales) | Pausa extendida o irregular | Reducir a espacio simple |
| Salto de Línea | ^l (flecha abajo) | Ruptura de prosodia / Sin pausa | Convertir a ^p |
| Guion Largo (—) | ^% (Em-dash) | Pausa parentética / Cambio de tono | Mantener (con espacios) |
| Tabulación | ^t (flecha derecha) | Silencio o error | Eliminar |
| Comillas Curvas | “ ” (Smart quotes) | Posible error de codificación | Normalizar a rectas " |

### 2.2 Gestión de la Puntuación y su Efecto Prosódico

La puntuación no es solo gramática; es la partitura musical para la voz neural. La red neuronal aprende de millones de muestras de audio humano y asocia signos de puntuación con patrones de respiración y entonación específicos.4

1. **La Coma (,):** No todas las comas son iguales. Una coma en una lista ("manzanas, peras, uvas") produce una entonación de lista (tono sostenido). Una coma vocativa ("Hola, Juan") produce una pausa breve y un cambio de foco. En la maquetación, evite el exceso de comas ("Oxford comma" opcional) si el flujo se siente entrecortado.
2. **El Punto y Coma (;):** Es una herramienta poderosa para el TTS. Genera una pausa más larga que una coma pero mantiene la conexión tonal entre las cláusulas, evitando la caída de tono final de un punto. Úselo para conectar ideas complejas sin romper el flujo narrativo.
3. **Puntos Suspensivos (...):** En la ficción, son vitales. Amazon Polly Neural interpreta ... excelentemente, introduciendo una pausa de duda o suspensión y, a menudo, ralentizando ligeramente las palabras precedentes para simular pensamiento. Asegúrese de usar el carácter de elipsis real (…) o tres puntos estándar consistentemente.20
4. **Guiones y Rayas:** La distinción es crucial.
   * *Guion corto (-):* Une palabras. "Bien-intencionado". El TTS lo lee fluido.
   * *Raya (Em-dash —):* Separa oraciones. "Vino hacia mí —con un cuchillo— y gritó". Aquí, la raya actúa como paréntesis dramático. Es vital rodear la raya de espacios delgados o normales. Si se pega a las palabras (mí—con), algunos motores intentarán leer "míguioncon" como una palabra compuesta extraña.18

## 3. Estructura Semántica: La Filosofía de los "Estilos" sobre el Formato Directo

El error fundamental en la creación de documentos para TTS es confiar en el formato visual directo (negrita, cursiva, tamaño de fuente). Un motor de conversión automatizada, como los basados en bibliotecas como mammoth.js o python-docx, ignora típicamente el formato directo a menos que se le indique explícitamente lo contrario. La solución profesional es el uso riguroso de **Estilos de Word**.22

### 3.1 Mapeo de Estilos a Comandos de Voz

En lugar de marcar un texto en cursiva para darle énfasis, debe aplicarse un estilo de carácter llamado "Énfasis". Esto permite que, en el proceso de conversión a audio, ese estilo específico se traduzca programáticamente a una etiqueta SSML <emphasis> o <prosody>.

#### 3.1.1 Jerarquía de Títulos (Headings) para la Navegación y Segmentación

La estructura de navegación del audiolibro depende de los estilos de título.

* **Heading 1 (Título 1):** Debe reservarse exclusivamente para los inicios de capítulo (ej. "Capítulo 1", "Prólogo").
  + *Función en Audio:* Este estilo sirve como "punto de corte" (split point). Los scripts de post-procesamiento pueden detectar Heading 1 para dividir el manuscrito en archivos de audio separados (Chapterization), un requisito obligatorio para plataformas de distribución como Audible/ACX.7
  + *SSML Implícito:* Al encontrar un Heading 1, el sistema puede inyectar automáticamente un silencio de 2 a 3 segundos antes de comenzar la lectura, asegurando que el oyente tenga un respiro mental entre capítulos.
* **Heading 2 y 3:** Utilizados para secciones dentro de un capítulo.
  + *Función en Audio:* Se traducen en pausas medias (1-2 segundos). No dividen el archivo, pero estructuran el ritmo.

#### 3.1.2 Estilos de Párrafo Personalizados

Para enriquecer la narración, se recomienda crear estilos personalizados en Word que no existen por defecto.

* **Estilo "Blockquote" (Cita en Bloque):** Para cartas, lecturas de documentos o sueños dentro de la novela.
  + *Efecto Neural:* Puede mapearse para activar un cambio sutil en la voz (ej. <prosody rate="slow">) o incluso cambiar el estilo de habla a "Lectura" si el motor lo soporta.
* **Estilo "Scene Break" (Salto de Escena):** A menudo representado visualmente por asteriscos (\*\*\*) o un espacio en blanco grande.
  + *Efecto Neural:* Debe mapearse a una etiqueta <break time="2s"/>. Si se deja solo como espacio visual en Word, el motor lo ignorará y saltará de una escena a otra sin pausa, confundiendo al oyente.4

### 3.2 Tabla de Mapeo Recomendada (Word a SSML)

A continuación se presenta una especificación técnica de cómo deben mapearse los estilos visuales a instrucciones SSML para un motor neural como Amazon Polly.10

| **Elemento Visual en Word** | **Estilo Recomendado** | **Etiqueta SSML/Acción (Post-proceso)** | **Efecto Psicoacústico** |
| --- | --- | --- | --- |
| Texto General | Normal | <p>...</p> | Voz base, ritmo neutro. |
| Título de Capítulo | Heading 1 | <break time="3s"/> + Nuevo Archivo | Pausa estructural mayor, reinicio de contexto. |
| Subtítulo | Heading 2 | <break time="1.5s"/> | Pausa de transición media. |
| Énfasis (Cursiva Visual) | Emphasis (Carácter) | <prosody volume="+2dB"> o <emphasis> | Aumento de intensidad y precisión articulatoria. |
| Pensamiento Interno | Internal (Párrafo) | <amazon:effect name="whispered"> | Efecto de intimidad/susurro (solo ciertas voces). |
| Diálogo Rápido | DialogueFast | <prosody rate="+10%"> | Aceleración para simular urgencia o discusión. |
| Separador de Escena (\*\*\*) | SceneBreak | <break time="2s"/> | Pausa narrativa clara sin cambio de archivo. |

**Nota Técnica:** Es vital verificar la compatibilidad de las etiquetas. Por ejemplo, la etiqueta <emphasis> estándar no es soportada por todas las voces neurales de Amazon Polly; a menudo se prefiere el control manual mediante <prosody> para garantizar resultados consistentes.25

## 4. Normalización de Texto (Text Normalization - TN): El Reto Oculto

La normalización de texto es el proceso de convertir representaciones simbólicas (números, fechas, abreviaturas) en su forma ortográfica completa. Aunque los motores TTS modernos tienen módulos de TN integrados, "adivinan" basándose en probabilidades, lo que lleva a errores frecuentes en contextos literarios o técnicos.28 Para un audiolibro profesional, la **normalización manual** en el archivo Word es obligatoria.

### 4.1 Desambiguación Numérica y Temporal

Los números son contextuales. El texto "1999" puede ser un año ("mil novecientos noventa y nueve" o "diecinueve noventa y nueve") o una cantidad ("mil novecientos noventa y nueve manzanas").

* **Estrategia:** No deje números como dígitos en el manuscrito a menos que sean fechas muy estándar. Escriba los números como palabras.
  + *Incorrecto:* "Llamó al 911 a las 5:00 PM."
  + *Correcto:* "Llamó al nueve uno uno a las cinco de la tarde."
  + *Razón:* Esto elimina cualquier posibilidad de que la IA lea "novecientos once" o "cinco cero cero p m".21

### 4.2 Acrónimos y Siglas

El dilema de "leer como palabra" vs. "deletrear".

* **Deletreo (Letter-by-letter):** Para siglas como "FBI" o "CIA", los motores suelen acertar, pero no siempre. Para forzar el deletreo en Word sin SSML, use espacios o puntos: "F.B.I.". Si usa SSML, la etiqueta <say-as interpret-as="spell-out"> es la correcta.
* **Lectura silábica (Acronyms):** Para "NASA" o "UNESCO", el motor debe leerlo como palabra. Si el motor falla y deletrea "N-A-S-A", la solución en Word es escribirlo fonéticamente o en minúsculas controladas si el motor es sensible a la caja (aunque esto es arriesgado). La mejor opción es usar <say-as interpret-as="cardinal"> no aplica aquí; se debe usar léxicos personalizados o sustitución fonética ("Nasa").15

### 4.3 Homógrafos y Heterónimos

Palabras con idéntica grafía pero diferente pronunciación y significado.

* *Ejemplo Inglés:* "Bass" (pez /bæs/) vs "Bass" (instrumento /beɪs/).
* Ejemplo Español: "Vino" (bebida) vs "Vino" (verbo).  
  Los motores neurales usan el contexto de la frase para resolver esto.32 Sin embargo, en poesía o frases cortas y ambiguas, pueden fallar. Si detecta un error de pronunciación constante en una palabra específica durante las pruebas, la solución es la Rescritura Fonética en el archivo Word para producción (ej. escribir "beis" en lugar de "bass" para el instrumento, si el idioma de síntesis lo permite, o usar IPA a través de SSML <phoneme>).33

## 5. Implementación Avanzada de SSML: El Código Detrás del Texto

El Speech Synthesis Markup Language (SSML) es el puente entre el texto plano y la interpretación rica. Aunque Word no soporta SSML nativamente, podemos usar el documento para "inyectar" o "marcar" dónde debe ir el código.

### 5.1 Inyección de Pausas y Ritmo (The <break> Tag)

El silencio es tan importante como el sonido. La falta de pausas adecuadas crea fatiga en el oyente.

* **Pausas Gramaticales vs. Dramáticas:** La coma automática genera una pausa breve. Para una pausa dramática ("Él abrió la puerta... y gritó"), los puntos suspensivos pueden no ser suficientes.
* **Técnica de Marcadores:** Inserte marcadores de texto plano en Word como o. Un script posterior de búsqueda y reemplazo (o una macro de Word) puede convertir `` en <break time="2s"/> antes de enviar el texto al motor.34

### 5.2 Control de la Respiración (Breathing)

La ausencia de sonidos de respiración es el marcador número uno de una voz sintética. Las voces neurales de Amazon Polly permiten la inserción de respiraciones artificiales que suenan realistas.

* **Automatización (<amazon:auto-breaths>):** Esta es una función exclusiva de ciertas voces de Polly. Permite que el motor inserte inhalaciones automáticamente basándose en la longitud de la frase.
  + *Implementación:* No se puede hacer a nivel de párrafo individual fácilmente en Word sin romper el flujo. Lo ideal es envolver bloques grandes de texto narrativa en esta etiqueta durante el post-procesamiento.35
* **Respiración Manual (<amazon:breath/>):** Para momentos de cansancio, miedo o ejercicio físico en la narración. Inserte un marcador `` en Word y conviértalo a <amazon:breath duration="medium" volume="loud"/> para simular un jadeo.36

### 5.3 Prosodia y Énfasis (<prosody> vs <emphasis>)

Aquí encontramos una limitación técnica importante. Muchas voces neurales de alta gama (NTTS) **ignoran** la etiqueta estándar <emphasis>. Para lograr énfasis, debemos manipular manualmente la prosodia.

* **Volumen y Velocidad:** Para enfatizar una palabra, la técnica más natural no es solo subir el volumen, sino **ralentizar** ligeramente la velocidad y aumentar el volumen.
  + *Fórmula de Énfasis:* <prosody rate="-10%" volume="+2dB">Palabra Importante</prosody>.
  + *En Word:* Cree un estilo de carácter "Énfasis Real" y configure su script de conversión (mammoth.js) para aplicar esta etiqueta compuesta compleja en lugar de una simple etiqueta <em>.26

## 6. Gestión de Diálogos y Multi-Voces (Casting Neural)

La narración de ficción presenta el desafío de distinguir personajes. Un audiolibro plano (una sola voz para todo) es aceptable, pero la tecnología actual permite producciones "multicast" sintéticas.

### 6.1 Estrategias de Asignación de Voces

Amazon Polly y otros servicios permiten cambiar el VoiceId dentro del mismo flujo de texto.

* **No use formato de guion:** Escribir "JUAN: Hola" es terrible para la experiencia auditiva si se lee tal cual.
* **Etiquetado Estilístico:** Asigne un estilo de carácter o párrafo único a cada personaje principal en Word (ej. Char\_Juan, Char\_Ana, Narrador).
  + El texto se ve así en Word: [Estilo Narrador] Él dijo: [Estilo Juan] "No quiero ir."
  + El script de conversión detecta el estilo y envuelve el texto: <voice name="Matthew">Él dijo:</voice> <voice name="Enrique">"No quiero ir."</voice>.38
* **Consideración de Costos y Latencia:** El uso de múltiples voces aumenta la complejidad del SSML y puede afectar los tiempos de renderizado y costos en algunos servicios, aunque en Polly se cobra por caracteres procesados independientemente de los cambios de voz.40

### 6.2 Estilos de Habla: Conversacional vs. Noticiero

Para el narrador, Amazon Polly ofrece variantes de estilo.

* **Conversational:** Ideal para la mayoría de la ficción moderna. Tiene una varianza de tono más "relajada".
* **Newscaster:** Ideal para citas de periódicos, noticias de radio dentro de la historia o libros de no ficción periodística.
* **Maquetación:** Use estilos de párrafo en Word (Style\_News) para marcar secciones que deban leerse con estilo de noticiero, mapeándolas a <amazon:domain name="news">.41

## 7. Elementos No Textuales: Imágenes, Tablas y Notas

Un manuscrito diseñado para impresión contiene elementos visuales que rompen la linealidad del audio.

### 7.1 Tablas: La Pesadilla del Audio

Las tablas no se pueden "leer" linealmente de manera coherente (Fila 1 Columna 1, Fila 1 Columna 2...).

* **Solución:** Las tablas deben ser **linealizadas** manualmente. El autor o editor debe reescribir la tabla como párrafos narrativos.
  + *Tabla:* | País | Capital |

| España | Madrid |

\* Texto Audio: "En el caso de España, la capital es Madrid."

* **Acción en Word:** Elimine la tabla física y reemplácela por el texto narrativo. Si el libro tiene ambos formatos (impreso y audio), mantenga una versión separada del archivo para audio ("Forking" del manuscrito).43

### 7.2 Notas al Pie y Referencias

Las notas al pie interrumpen el flujo. En un audiolibro, no se puede "mirar abajo" y volver.

* **Ficción:** Elimine las notas al pie o intégrelas en el texto si son esenciales.
* **No Ficción:** Agrupe las notas al final del capítulo o del libro como una sección separada ("Notas del Capítulo 1"). No use hipervínculos ni marcadores de nota automática de Word, ya que el TTS podría leer el número de la nota ("uno") en medio de una frase.45

### 7.3 Texto Alternativo (Alt Text) para Imágenes

Si una imagen es crucial, escriba una descripción explícita en el cuerpo del texto con un estilo ImageDesc. No confíe en el atributo "Alt Text" de las propiedades de imagen de Word, ya que la mayoría de los conversores a texto plano lo ignoran. Escriba: "".8

## 8. Flujo de Trabajo Técnico Automatizado: De Word a Audio

Para una producción escalable, no se debe copiar y pegar texto manualmente. Se recomienda un flujo de trabajo basado en código.

### 8.1 Extracción con Mammoth.js

mammoth.js es una biblioteca fundamental que convierte .docx a HTML basándose en estilos semánticos, descartando la basura visual.22

Configuración del Style Map (Mapeo de Estilos):

Debe crear un archivo de configuración que diga a Mammoth cómo interpretar sus estilos de Word.

JavaScript

// Ejemplo conceptual de mapeo en Mammoth.js  
var options = {  
 styleMap: => h1:fresh", // Para dividir capítulos  
 "p[style-name='Cita'] => blockquote:fresh", // Para cambios de voz  
 "r[style-name='Emphasis'] => em", // Énfasis básico  
 "r => strong" // Énfasis fuerte  
 ]  
};

.10

### 8.2 Script de Python para Segmentación y SSML

Una vez obtenido el HTML limpio, un script de Python puede:

1. Parsear el HTML.
2. Dividir el contenido cada vez que encuentre un <h1> (Capítulo), creando archivos individuales.
3. Reemplazar las etiquetas HTML con etiquetas SSML específicas de Polly (ej. <em> a <prosody>).
4. Añadir el "boilerplate" necesario: <speak>, <amazon:auto-breaths>, y los silencios de sala requeridos.48

## 9. Cumplimiento de Estándares de Distribución (ACX/Audible)

La maquetación final debe asegurar que el audio generado cumpla con los requisitos técnicos de las plataformas de distribución.

### 9.1 Silencios de Sala (Room Tone)

ACX exige un silencio al inicio (0.5 - 1 seg) y al final (1 - 5 seg) de cada archivo. El TTS puro termina abruptamente.

* **Solución:** Inyecte <break time="3s"/> al final de cada capítulo mediante su script de conversión. No intente hacerlo con espacios en blanco en Word, ya que son impredecibles.7

### 9.2 Niveles de Audio (RMS y Picos)

Aunque esto es post-producción de audio, el archivo Word influye. Si usa demasiadas etiquetas <prosody volume="loud">, puede causar distorsión (clipping) o violar el límite de -3dB de pico.

* **Recomendación:** Use cambios de volumen relativos moderados (+2dB o loud, nunca x-loud masivamente) para mantener el rango dinámico bajo control y facilitar la masterización posterior.40

### 9.3 Limitaciones Legales

Es fundamental recordar que, a fecha de este informe, ACX/Audible **no permite** audiolibros narrados completamente por IA para la mayoría de los usuarios, exigiendo narración humana. Sin embargo, este flujo de trabajo es válido para:

1. Producción de maquetas (proof-of-concept).
2. Plataformas que sí permiten IA (Google Play Books, Kobo, venta directa).
3. Contenido accesible para discapacitados visuales.50

## 10. Lista de Verificación de Maquetación (Checklist)

Antes de procesar su archivo Word, verifique:

1. **Limpieza:** ¿Cero tabulaciones, cero dobles espacios, cero saltos de línea manuales (^l)?
2. **Estilos:** ¿Todo el texto tiene estilo Normal o Heading? ¿Se ha eliminado todo formato directo (negrita/cursiva manual)?
3. **Semántica:** ¿Los títulos de capítulo son Heading 1? ¿Los cambios de escena tienen su propio estilo o marcador?
4. **Normalización:** ¿Números, fechas y siglas están expandidos a texto completo?
5. **Tablas/Imágenes:** ¿Se han eliminado y reemplazado por descripciones narrativas?
6. **Caracteres Especiales:** ¿Se han revisado las rayas (—) y comillas para asegurar la codificación correcta?

## 11. Conclusión

La creación de un audiolibro con voz neural de alta calidad no es una tarea de "convertir y listo". Es un proceso de **remediación de contenido**. El archivo Word debe dejar de verse como un documento visual para convertirse en una base de datos estructurada de instrucciones prosódicas. Al aplicar una higiene tipográfica estricta, utilizar estilos semánticos para controlar el SSML y normalizar manualmente el texto, los creadores pueden superar las limitaciones robóticas de la IA y producir obras auditivas que respiran, sienten y narran con una naturalidad sorprendente. La "magia" de la voz neural no está solo en el algoritmo, sino en la calidad de los datos que le alimentamos.

#### Obras citadas

1. Neural voices - Amazon Polly - AWS Documentation, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/polly/latest/dg/neural-voices.html>
2. Text-to-Speech 101: The Ultimate Guide | by Felix Laumann, PhD | NeuralSpace | Medium, fecha de acceso: diciembre 15, 2025, <https://medium.com/neuralspace/text-to-speech-101-the-ultimate-guide-9a4b10e20fef>
3. AI Narration in Audiobooks: A Game-Changer for Authors - Spines, fecha de acceso: diciembre 15, 2025, <https://spines.com/ai-narration-in-audiobooks-a-game-changer-for-authors/>
4. How do TTS systems handle punctuation and formatting cues? - Milvus, fecha de acceso: diciembre 15, 2025, <https://milvus.io/ai-quick-reference/how-do-tts-systems-handle-punctuation-and-formatting-cues>
5. More-natural prosody for synthesized speech - Amazon Science, fecha de acceso: diciembre 15, 2025, <https://www.amazon.science/blog/more-natural-prosody-for-synthesized-speech>
6. Improving the naturalness of an end-to-end text-to-speech system with information structure, fecha de acceso: diciembre 15, 2025, <https://campus-fryslan.studenttheses.ub.rug.nl/235/1/MA%204954475%20J%20Baumann.pdf>
7. ACX audio submission requirements - ACX Help Center, fecha de acceso: diciembre 15, 2025, <https://help.acx.com/s/article/what-are-the-acx-audio-submission-requirements>
8. Best practices for Creating Accessible Microsoft Office Documents, fecha de acceso: diciembre 15, 2025, <https://accessibility.calpoly.edu/content/instmaterials/document_creation/best_practices>
9. VisualSpeech: Enhance Prosody with Visual Context in TTS - arXiv, fecha de acceso: diciembre 15, 2025, <https://arxiv.org/html/2501.19258v1>
10. mammoth.js/README.md at master · mwilliamson/mammoth.js · GitHub, fecha de acceso: diciembre 15, 2025, <https://github.com/mwilliamson/mammoth.js/blob/master/README.md>
11. Text-to-Speech Best Practices for Internal and Public Alerts - CivicPlus Help Center, fecha de acceso: diciembre 15, 2025, <https://www.civicplus.help/mass-notification/docs/text-to-speech-best-practices-for-internal-and-public-alerts>
12. Remove Extra Spaces and Double Spaces - PhraseFix, fecha de acceso: diciembre 15, 2025, <https://phrasefix.com/tools/remove-extra-spaces/>
13. Cleaning Text with python and re - Stack Overflow, fecha de acceso: diciembre 15, 2025, <https://stackoverflow.com/questions/55187374/cleaning-text-with-python-and-re>
14. Part 3: Step by Step Guide to NLP - Text Cleaning and Preprocessing - Analytics Vidhya, fecha de acceso: diciembre 15, 2025, <https://www.analyticsvidhya.com/blog/2021/06/part-3-step-by-step-guide-to-nlp-text-cleaning-and-preprocessing/>
15. SSML tags supported by Amazon Connect - AWS Documentation, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/connect/latest/adminguide/supported-ssml-tags.html>
16. Whitespace Cleanup - Remove Extra Spaces, Tabs, Line Breaks in Your Text - Strip HTML, fecha de acceso: diciembre 15, 2025, <https://www.striphtml.com/text-tools/whitespace-cleanup/>
17. Lessons learned from building an AI-powered file parser for document cleaning - Reddit, fecha de acceso: diciembre 15, 2025, <https://www.reddit.com/r/software/comments/1oc9pls/lessons_learned_from_building_an_aipowered_file/>
18. Cleaning AI Generated Text | Greg's Half Ideas, fecha de acceso: diciembre 15, 2025, <https://gregraiz.com/blog/cleaning-ai-generated-text/>
19. Speech Synthesis Markup Language (SSML) - Google Cloud Documentation, fecha de acceso: diciembre 15, 2025, <https://docs.cloud.google.com/text-to-speech/docs/ssml>
20. Add pauses to text-to-speech voiceovers - Narakeet, fecha de acceso: diciembre 15, 2025, <https://www.narakeet.com/docs/how-to/add-pauses-to-text-to-speech-voiceovers.html>
21. Text to Speech - Best Practices - Waves, fecha de acceso: diciembre 15, 2025, <https://waves-docs.smallest.ai/v3.0.1/content/best-practices/tts-best-practices>
22. mwilliamson/mammoth.js: Convert Word documents (.docx files) to HTML - GitHub, fecha de acceso: diciembre 15, 2025, <https://github.com/mwilliamson/mammoth.js/>
23. Mammoth: Convert Word documents from docx to simp Guide - Generalist Programmer, fecha de acceso: diciembre 15, 2025, <https://generalistprogrammer.com/tutorials/mammoth-npm-package-guide>
24. Splitting a docx by headings into separate files in Python - Stack Overflow, fecha de acceso: diciembre 15, 2025, <https://stackoverflow.com/questions/54409495/splitting-a-docx-by-headings-into-separate-files-in-python>
25. Supported SSML tags - Amazon Polly - AWS Documentation, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html>
26. Controlling volume, speaking rate, and pitch - Amazon Polly - AWS Documentation, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/polly/latest/dg/prosody-tag.html>
27. Emphasis and neural voices | Articulate - Community - E-Learning Heroes, fecha de acceso: diciembre 15, 2025, <https://community.articulate.com/discussions/discuss/emphasis-and-neural-voices/913695>
28. Benefits - Amazon Polly - AWS Documentation, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/polly/latest/dg/benefits.html>
29. Text Normalization for Text-to-Speech - uu .diva, fecha de acceso: diciembre 15, 2025, <https://uu.diva-portal.org/smash/get/diva2:1764605/FULLTEXT01.pdf>
30. TTS1 English Text Normalization | LumenVox Knowledgebase, fecha de acceso: diciembre 15, 2025, <https://help.lumenvox.com/knowledgebase/index.php?/article/AA-01886/0/TTS1-English-Text-Normalization.html>
31. Controlling how special types of words are spoken - Amazon Polly, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/polly/latest/dg/say-as-tag.html>
32. Amazon Polly – Text to Speech in 47 Voices and 24 Languages | AWS News Blog, fecha de acceso: diciembre 15, 2025, <https://aws.amazon.com/blogs/aws/polly-text-to-speech-in-47-voices-and-24-languages/>
33. Customize pronunciations using Amazon Polly | Artificial Intelligence - AWS, fecha de acceso: diciembre 15, 2025, <https://aws.amazon.com/blogs/machine-learning/customize-pronunciations-using-amazon-polly/>
34. Adding a pause - Amazon Polly - AWS Documentation, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/polly/latest/dg/break-tag.html>
35. Adding the sound of breathing - Amazon Polly, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/polly/latest/dg/breath-tag.html>
36. Amazon Polly releases new SSML Breath feature | Artificial Intelligence - AWS, fecha de acceso: diciembre 15, 2025, <https://aws.amazon.com/blogs/machine-learning/amazon-polly-releases-new-ssml-breath-feature/>
37. How to Convert DOCX To Html With Python Mammoth - Stack Abuse, fecha de acceso: diciembre 15, 2025, <https://stackabuse.com/how-to-convert-docx-to-html-with-python-mammoth/>
38. Speech Synthesis Markup Language (SSML) Reference | Alexa Skills Kit, fecha de acceso: diciembre 15, 2025, <https://developer.amazon.com/en-US/docs/alexa/custom-skills/speech-synthesis-markup-language-ssml-reference.html>
39. Question for narrators about recording multivoiced audiobooks : r/ACX - Reddit, fecha de acceso: diciembre 15, 2025, <https://www.reddit.com/r/ACX/comments/s3almq/question_for_narrators_about_recording/>
40. Amazon Polly Pricing Guide: Cost Optimization & Savings Strategies - Astuto AI, fecha de acceso: diciembre 15, 2025, <https://www.astuto.ai/blogs/amazon-polly-pricing-and-cost-optimization>
41. Applying the newscaster voice - Amazon Polly - AWS Documentation, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/polly/latest/dg/newscaster-voices.html>
42. Engage listeners with Amazon Polly's Conversational speaking style voices - AWS, fecha de acceso: diciembre 15, 2025, <https://aws.amazon.com/blogs/machine-learning/engage-listeners-with-amazon-pollys-conversational-speaking-style-voices/>
43. How to remove table in MS Word while retaining the text inside of it - Super User, fecha de acceso: diciembre 15, 2025, <https://superuser.com/questions/191775/how-to-remove-table-in-ms-word-while-retaining-the-text-inside-of-it>
44. Remove Table in MS Word without Deleting Text | Delete Table but Keep Text - YouTube, fecha de acceso: diciembre 15, 2025, <https://www.youtube.com/watch?v=IwhOVY4rCfo>
45. Accessible Audiobook Workflow Guide: Producing Born Accessible Audiobooks - APLN, fecha de acceso: diciembre 15, 2025, <https://apln.ca/accessible-audiobook-workflow-guide-producing-born-accessible-audiobooks/>
46. AC-T-36 Handling Footnotes + Endnotes In Nonfiction Audiobooks - Pro Audio Voices, fecha de acceso: diciembre 15, 2025, <https://proaudiovoices.com/audiobook-production-tip-footnotes-and-endnotes-36/>
47. 2 Adding Alt Text to Images & Tables - Georgia Digital Service, fecha de acceso: diciembre 15, 2025, <https://digital.georgia.gov/2-adding-alt-text-images-tables>
48. Generating speech from SSML documents - Amazon Polly, fecha de acceso: diciembre 15, 2025, <https://docs.aws.amazon.com/polly/latest/dg/ssml.html>
49. Audiobook Mastering - Audacity Support, fecha de acceso: diciembre 15, 2025, <https://support.audacityteam.org/audio-editing/audiobook-mastering>
50. ACX doesn't allow AI generated voices. It's in their terms of service. They will automatically reject these submissions. I know, I've tried. I can't speak for the other platforms, but this is… - Jason Lehigh - Medium, fecha de acceso: diciembre 15, 2025, <https://medium.com/@jlehigh/acx-doesnt-allow-ai-generated-voices-559cb9b45df6>
51. Can You Use AI Voice for Audible, ACX, and Findaway Voices? - Narration Box, fecha de acceso: diciembre 15, 2025, <https://narrationbox.com/blog/can-you-use-ai-voice-for-audible-acx-and-findaway-voices-(2025-guide)>