Archivum, LXXIV, 2024, pp. 577-617

# The deaccenting of given information in English in TTS systems: a case study

Alfonso Carlos Rodríguez Fernández-Peña
Universidad de Oviedo
rodriguezalfonso@uniovi.es

Recibido: 30/08/2023
Aceptado: 15/04/2024

**ABSTRACT:**
*This paper provides a descriptive qualitative and quantitative study of the deaccenting of given information, a.k.a.* anaphora rule, *by four well-known online Text-to-Speech (TTS) software (Murf, Lovo, PlayHT, and Replica Studios). We have used 10 sentences as input, each containing elements of given information to test the software. The voice types selected for our analysis are one male with a British English accent and one female with an American English accent for each software. Eachsentence has been uttered by the voice skins in each software, downloaded in audio format, and analysed using the speech analysis software Praat. This way we can measure and evaluate the pitch contours for each utterance and check whether the anaphora rule is applied or not by the different TTS software. The general results show that almost 70% of the lines do not achieve the delivery of the anaphora rule. This means that this prosodic feature characteristic of English stress and the substantial pragmatic load it carries is lost most of the time. The results obtained indicate that even though synthetic voices may be successful at the segmental level in terms of catenation and voice quality, the suprasegmentals and prosodic elements of human speech are not mastered by the machines yet.*

**KEYWORDS:** *Text-to-speech; prosody; deaccenting; anaphora rule; synthetic voices; tonicity*

# La desacentuación de la información conocida en inglés en los sistemas TTS: un estudio de caso

**RESUMEN:**
Este artículo ofrece un estudio descriptivo tanto cualitativo como cuantitativo de la desacentuación de la información conocida en inglés por cuatro programas de conversión de texto a voz (TTS) en línea (Murf, Lovo, PlayHT y Replica Studios). Como texto de entrada para probar estos programas se han utilizado diez frases en las que cada una conteniene elementos de información conocida. Los tipos de voces inglesas seleccionadas para nuestro análisis son una voz masculina con acento británico y una voz femenina con acento estadounidense para cada software. Cada oración ha sido reproducida por las voces en cada software, descargada en formato de audio y analizada utilizando el software de análisis acústico Praat. De esta manera, hemos medido y evaluado los contornos tonales para cada enunciado y comprobado si la regla de la anáfora se aplica o no en los diferentes programas TTS. Los resultados generales muestran que casi el 70 % de las oraciones reproducidas por estos programas no logran aplicar la regla de la anáfora, lo que significa que esta característica prosódica propia del inglés y su correspondiente carga pragmática se pierde la mayoría de las veces. Los resultados obtenidos indican que, aunque las voces sintéticas pueden ser exitosas a nivel de producción segmental en términos de concatenación y calidad de voz, los elementos suprasegmentales y prosódicos del habla humana aún no son del todo reproducibles por las máquinas.
**PALABRAS CLAVE***: Texto a voz; prosodia; desacentuación; regla de la anáfora; voces sintéticas; tonicidad*

## 1. Introduction

Synthetic voices are in vogue. We live in a world in which the creation of audiovisual content is massive and in continuous increase. Artificial intelligence companies are starting to bring celebrities back to life for commercial purposes, and living celebrities are giving their vocal and physical rights to these companies to be exploited in the future, in what has been coined as *deepfakes*. Synthetic voices are a reality and the number of com-

panies offering text-to-speech services is soaring online. This paper aims to analyse four online providers of text-to-speech services and whether their voices apply one prosodic phenomenon called the *anaphora rule*, or the deaccenting of given information. We have selected two voices (one male British and one female American) for each of the providers and have had them deliver 10 sentences in which any English native speaker would apply this human prosodic feature. First, we will comment on TTS and voice-speaking software, trying to understand what it is and how it works. Then, we will delve into how prosody is achieved and conveyed in TTS, followed by an explanation of how English tonicity works and how the anaphora rule is conveyed. The objectives of the study, the corpus selected, and the methodology used will be commented just next. Subsequently, the results obtained will be displayed together with the analysis of the utterances that successfully applied the delivery of the anaphora rule. Finally, we will offer a conclusion and new possible lines of research to understand how prosodic traits are conveyed by these voice skins and the way these can become game changers in the voice and communication industry.

**2. TTS and voice-speaking software**

Text-to-speech has been defined by Dutoit (1997a: 13) as "the production of speech by machines, by way of the automatic phonetization of the sentences to utter". The ultimate objective of a text-to-speech synthesizer should be to read any text as naturally and comprehensively as possible, regardless of whether it was entered directly into the computer by an operator or scanned and submitted to an optical character recognition (OCR) system.

The way a common text-to-speech system works has been widely discussed and explained by scholars such as Dutoit (1997a, 1997b), Taylor (2009), Hassid et al. (2022), or Tan et al. (2022), among others, and can be explained as follows. The input text arrives as an arbitrary-length string of ASCII charac-

ters. To make processing more manageable, we split the input text into discrete sentences using an algorithm for sentence splitting. We do not know whether the input contains only one sentence, thus we always attempt to identify sentence boundaries. Based on the existence of whitespace, punctuation, etc., we divide the input into a series of tokens for each sentence. Typically, tokens are written encodings of individual words, but they can also be encodings of integers, dates, and other data kinds. Next, each token's semiotic class is determined. For non-natural language tokens, a distinct method is used for each kind to decode the text into the underlying form, and then a set of rules is applied to convert this form into a natural language form containing words. We seek to resolve any ambiguity in natural language tokens to locate their words. A basic prosodic analysis of the text is attempted using algorithms to determine the utterance's phrasing, emphasis patterns, and intonation, although the text lacks a substantial amount of the information we would ideally like. At this point, the text and prosodic analysis phase has concluded. The initial step in the synthesis process is to encode the newly discovered words as phonemes. This is done to offer a more compact representation for subsequent synthesis processes to operate on. The words, phonemes, and phrase structure constitute an input specification for the unit selection module. Actual synthesis is accomplished by searching a database of pre-recorded speech for units that fit the input specification as nearly as feasible. The prerecorded speech can be stored as a database of waveform fragments; when a particular sequence of these fragments is selected, signal processing is employed to stitch them into a single continuous output speech waveform. Essentially, this is how TTS operates.

The diagram in Figure 1, from Dutoit 1997a, provides an overall view of how a standard TTS process develops.

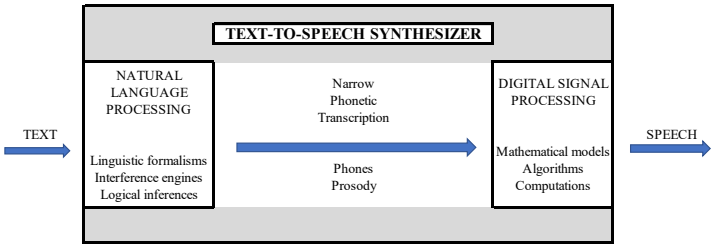



Figure 1. A functional diagram of a TTS system (Dutoit 1997a: 14)

The model represented in Figure 1 by Dutoit (1997) belongs to what Taylor (2009) labelled as the *common form model,* for which there are essentially two components: "a text analysis system that decodes the text signal and uncovers the form, and a speech synthesis system which encodes this form as speech" (2009: 37). Nonetheless, there are also other TTS models, as explained by Taylor (2009), not mutually exclusive, which can be combined to make real life TTS systems. From these, the most relevant in our study are:

Grapheme and phoneme form models: This method parallels the common form model in that a grapheme form of the input text is discovered before translation into a phoneme form for synthesis. In contrast to the usual form paradigm, the focus is not on words. This strategy is particularly helpful for languages in which the correspondence between grapheme and phoneme is fairly straightforward; in such languages, discovering the graphemes generally involves locating the phonemes and, consequently, the right pronunciation.

Complete prosody generation: The three distinct forms of prosody (affective, suprasegmental, and augmentative —as seen in section 3—) can be described separately using the common form model. The verbal part serves as a model for suprasegmental phonetics, and this is true even for statements with less affective prosody. As a result, only augmentative prosody needs to be developed consciously. This goes against the belief held by

many systems that the F0, phrasing, stress, etc. of an utterance are directly dictated by prosody, and that it is necessary to create all these quantities with an explicit prosodic model to make natural-sounding speech. The prosodic component of the system would be more important if this were the case.

Prosody from the text: If every utterance needs a comprehensive prosodic specification, as the complete prosody model assumes, then this must be generated. The text is assumed to have sufficient information to establish prosody, leading to the inclusion of modules in many TTS systems that attempt to anticipate prosodic representations directly from the text.

Apart from the TTS models described by Taylor (2009), more recent ones have been developed, like Tahon et al.'s (2017) phoneme-to-phoneme model (P2P), which blends the grapheme-to-phoneme model described by Taylor (2009) and conditional random fields (CRFs), which consist of statistical modelling methods applied in pattern recognition and machine learning for structured prediction. This model combines linguistic features with articulatory and prosodic ones; and, according to objective and perceptual tests (Qader et al. 2017), it reflects pronunciation and spontaneous speech better than non-adapted models.

***2. 1 Other speech technologies***

Text-to-speech systems must compete with other types of products that produce *canned* messages from waveform-coding chips, as previewed by Klatt (1987) back in the 1980s. These pre-recorded prompts imply, as explained by Taylor (2009: 43), that a designer creates an exhaustive list of utterances that are needed for an application. Once these lines or utterances are designed, a voice talent is asked to read them, they are recorded, and stored in digital audio format. These kinds of voice message systems are commonly known as *voice user interfaces* (VUI) and unlike TTS, these are unique in that they are based on spoken language (Cohen et al. 2004). A VUI is part of a spoken language application that a person interacts with when speaking

with the programme. Prompts, grammars, and dialogue logic are some of the components that make up a VUI. During the conversation between the user and the system, the user will hear recordings or synthetic speech that are referred to as prompts. A VUI is the kind of system used by certain companies in their support or customer service. Here, a voice talent has prerecorded everything the system has to say. Following a prompt such as "Are you calling because of a problem with your computer or your phone?", the system listens searching for inputs from the caller such as "computer" or "phone", "smartphone", or "telephone". Then, the dialog logic decides what to do afterward depending on the answer, and puts the listener through to a human specialist or continues with more prompts until there is a final solution for the client's needs. Amongst the most well-known VUI are Apple's *Siri*, Microsoft's *Cortana*, and Amazon's *Alexa*.

Another speech technology that has spread in recent years is speaker verification technology, which helps ensure that the person on the other end of a phone conversation is who they say they are. It has been implemented for a wide variety of applications, and one use of it has been as a component of the authentication procedure for a spoken language application (Cohen et al. 2004). In several applications, personal identification numbers (PINs), which consumers previously needed to memorise, have been replaced by speaker verification so that customers no longer need to remember them. This technology became popular in the 1990s thanks to the film *Sneakers* (1992).

**3. TTS and prosody**

There is no straightforward definition for the term *prosody* since it involves different aspects of oral language. Cruttenden (2014: 04) understands prosody as "how words and sentences are accented, and how pitch, loudness, and length work to produce rhythm and intonation", a definition which is shared by Ashby (2011: 141). Dutoit (1997: 129) considers that prosody re-

fers to the properties of the speech signal that produce audible changes such as pitch, loudness, and syllable length, and points that intonation is sometimes used as a synonym for prosody. In addition, this scholar states that, for some authors, prosodic features also include rhythm and speech rate, as in the case of Cruttenden (2014). Moreover, Cohen et al. (2004: 117) believe that prosody is the nonverbal meaning of language, that it is "an element of grammar, the implicit knowledge that native speakers have about their language". All these definitions of prosody show that there are no widely agreed description systems for any aspect of prosody. For this reason, as noted by Taylor (2009), we consider it worth alerting the reader to the extremely tenuous literature concerning prosody.

According to Taylor (2009), there are three major types of prosody, which facilitate the generation of prosody in a text-to-speech system: affective prosody, suprasegmentals, and augmentative prosody. Affective prosody, or "pure prosody" (Taylor 2009: 123) encompasses the expression of emotional, mental, and speaker attitude-related meaning. Significantly, prosody is the principal conduit through which these meanings are communicated. There is a considerable degree of linguistic universality for affective prosody; in all languages, raising one's voice and yelling are considered indications of anger or hostility, whereas speaking in a quieter voice is considered conciliatory or soothing. Besides emotion, affective prosody also involves the speaker attitude.

Suprasegmental elements include, following Taylor's (2009) model, intonational patterns, pitch variation, or speech melody, which are not included under the umbrella of prosody since these belong to what this scholar refers to as "verbal phonetics" (2009: 125). This is evident, in Taylor's view, because speakers do not communicate more information through the employment of these effects, and they are as innate as, for example, nasalization, voicing, and tongue position. This poses a problem in tone languages like Mandarin Chinese, which uses pitch to identify

words. Nonetheless, although intonation and sections of word identity are conveyed in the same acoustic variable, which complicates the analysis and synthesis, recent synthesis techniques allow us to model both without difficulty.

A message's verbal component is successfully communicated with the use of augmentative prosody. This type, unlike affective prosody, does not include or convey any additional information; it is simply a means for the speaker to ensure that a message is decoded and comprehended more precisely. The way augmentative prosody works is by modifying the suprasegmental default content of an utterance in specified ways. It is employed primarily to increase the likelihood that the verbal message will be successfully decoded and understood. Notably, unlike affective prosody, it does not add any additional information to the message.

### *3.1 TTS Prominence Prediction*

The basic approaches for prominence prediction algorithms used in TTS are the same as for phrase break prediction, where, as explained by Koutny et al. (2000) and Taylor (2009), there are simple deterministic algorithms, sophisticated deterministic algorithms, and data-driven algorithms. Phrasing is a key topic in the linguistic component of text-to-speech technologies and, according to Agüero and Bonafonte (2003: 107), consists of the process of dividing long sentences into smaller prosodic phrases. Acoustically, boundaries are defined by a pause, a tonal shift, and a lengthening of the final syllable. Punctuation is closely related to prosody, and —in many cases—, its primary function is related to syntax rather than acoustics. Phrase breaks have a significant impact on a sentence's naturalness, intelligibility, and interpretation, and their presence or absence can alter the meaning of a sentence.

Taylor (2009: 137) explains that the simplest prominence prediction algorithm uses the concatenation of the lexical prominence patterns of words, as illustrated below:

(1) In terms of the necessary political expediency to ensure survival – the result was a clearly a good one for Saw.

The prominence pattern we could generate is shown in example 2, where the prominent syllables are highlighted in bold:

(2) In **terms** of the **ne**cessary po**li**tical ex**pe**diency to en**sure** sur**vi**val – the re**sult** was a **clea**rly a **good one** for **Saw**.

There have been significant improvements in the naturalness of synthesized speech, due in great part to the refinement of concatenative synthesis. Cohen et al. (2004: 24) explain that a concatenative synthesizer utilizes a vast library of recorded speech samples. A succession of these prerecorded segments is concatenated to form the output signal. Signal processing is used to establish the proper timing and intonation contour and to smooth down the segment boundaries so that concatenation splices are inaudible.

**4. Tonicity: the focusing function of English intonation**

The British School proposes an intonational model based on two configurations, the nuclear configuration and the pre-nuclear configuration (Estebas-Vilaplana 2014), and suggests dividing each intonation phrase, or IP, into the *pre-head*, *head*, *nucleus*, and *tail* components. IPs can be composed of a single word (*yes!*) or multiple words (*Can you help me?*). The only rule that applies to all IPs is that they must have a single intonation nucleus, which is the most prominent syllable; it "typically has a marked change in pitch, and is somewhat longer and louder than the rest" (Collins and Mees 2013: 142).

The nucleus is used to mark the pertinent information in each IP, typically the most recent or new. The pitch pattern carried by the nucleus is the *nuclear tone*, which begins and ends on this syllable if it is the final syllable in the IP. Otherwise, the tone will be completed throughout the syllables which follow the nucleus, which are known as the *tail* of the IP. The elements occurring before the nucleus within an IP are referred to as the *head* and

*prehead*; the former comprises the elements within an IP from the first accented syllable up to and including the syllable preceding the nucleus, while the *prehead* comprises any unaccented syllables preceding the *head*.

| Prehead | Head | Nucleus | Tail |
|---|---|---|---|
| But | ˈcouldn't we ˈleave it till | ˈFri | day? |

Table 1. Structure of an IP according to the British School.

Amongst the main functions of English intonation, focusing is achieved through *tonicity* (the positioning of the nucleus) and the placement of other accents. When people speak, they produce IPs that may contain one or more words and not all of the words within the IPs have the same significance. Within each IP, native speakers select one word as especially significant for meaning, typically carrying new information, and place the nucleus, or most prominent accent, which will carry the nuclear tone or final pitch movement, on that word. In English, the nucleus is typically located on the final word of the IP's content. Tonicity, then, is typically determined by whether the words in an utterance convey new information or not. Let us analyse the sentence: *Meet me in front of the pub at* ***s****even*.

The natural location for the tonic syllable (the nucleus) is the first syllable of the word "seven", the final content word in the IP. In this instance, the nucleus denotes the conclusion of new information. The speaker is emphasizing the entire plan to meet in front of the pub at seven, not just the number seven. In this example, the focus is *broad* and encompasses the entire clause. When the nucleus or tonic syllable falls on the final lexical item of the IP, the tonicity or position of the nucleus within the IP is neutral, i.e., *neutral tonicity* (Tench 2009: 56).

However, in some instances, only a portion of the information in an utterance is highlighted. This is referred to as a *narrow* focus, and typically, old or given information is left out of focus.

Consider the example below from Wells (2006: 118) to see how narrow focus works:

| | |
|---|---|
| Who's bringing the food? | 'Mary. |
| | 'Mary is. |
| | 'Mary's bringing it. |
| | 'Mary's bringing the food. |
| | It's 'Mary that's bringing it. |

There may be cases in which the conversational context compels the speaker to alter the tonic syllable and the emphasis of an utterance. Let us see how this may work with the following dialogue.

A- Meet me in front of the pub at **se**ven.
B- OK, I'll see you in front of the theatre at seven.
A- No. Meet me in front of the **pub** at seven.
B- OK, I'll see you in the pub at seven.
A- No. Meet me in **front** of the pub at seven.

In the examples above, the emphasis is placed not on the final word of the content, as in broad focus, but on specific words that the speaker wishes to emphasize and bring into focus. This is known as *contrastive* focus and is a type of *narrow focus* because the utterance contains both new and previous information. These are examples of *marked* tonicity, which typically occurs when the nucleus, i.e., the tonic syllable, does not fall on the very last lexical item of the IP, its final word. There are, however, instances of narrow focus with neutral tonality in which the nucleus falls on the very last word of the IP. This is the case when all the information preceding the nucleus is outdated or already known. Tench (2009: 59) illustrates this with the following example:

A- I think I'll go and have a cup of **tea**.
B- (Well) why don't you come and have a spot of **lunch**?

As can be seen, B shows narrow focus and neutral tonality. This can be explained from the context, since "a spot of lunch" is the only new piece of information; "you come and have" is a reference to "I'll go and have" from the previous IP in A. Given that the nucleus should denote the conclusion of new information, "lunch" should carry the nucleus in this instance.

**5. The anaphora rule and the deaccenting of given information**

Among the causes that may hinder the development of effective models of prosody is that scholars have frequently attempted to investigate prosody without regard to its communicative role (Taylor 2009: 123). It is extremely typical in the field of verbal linguistics to examine a given sentence without regard to why a speaker uttered it, and this separation generally provides useful modularity for research purposes. However, prosody is informative. It draws attention to what is new and discloses the speaker's belief, intent, and understanding in ways that we take for granted. Therefore, the pragmatic load of utterances plays a substantial, even crucial, role in oral communication and should be carefully considered in successful TTS systems.

For J. L. Austin (1962) (Hatim & Mason 1990: 59), the pragmatic dimension of every *speech act*[1] consists of three distinct actions: locutionary act, illocutionary act, and perlocutionary act.

The illocutionary force implied in every speech act is the "real driving force of communication" (Mateo 2014: 125) and, consequently, the way it is conveyed through prosody should be highly considered.

In English, old information, information that has already been presented, or information that is repeated, is destressed and consequently *deaccented* (Halliday 1967, Prince 1981, Hirschberg 2006, Wells 2006). Therefore, the placement of the nucleus "signals the end of the new information in an intonation phrase"

[1] Actions performed via utterances are generally called speech acts according to Yule (1996: 47).

(Wells 2006: 109). This criterion also applies to any repeated terms or near-synonyms in a dialogue, as this signifies that the information is already known, assumed, and hence deaccented. Mott (2011: 205) refers to this prosodic phenomenon as the "anaphora rule".

Wells (2006: 109) exemplifies this prosodic effect with the following examples:

(3) How about a gin and tonic? - Oh, I'd prefer a **vod**ka and tonic.
(4) D'you object to dogs? - No, I a**dore** dogs.
(5) Who doesn't want to dance? - **Bill** doesn't want to dance.

And Mott (2011: 205):

(6) How many times did he hit you? - **Three** times.
(7) Dave and Jill decided to buy a **dog** and, when they'd **bought** the dog, they didn't want it in the house.

As can be seen in example 3 *tonic* has already been mentioned and, consequently, is old information. Therefore, it is deaccented in the answer and the nucleus falls on the first syllable of *vodka*. The same applies to *dogs* in example 4, to *doesn't want to dance* in example 5, and to *times* in example 6. In example 7 we can see that the second *dog* is deaccented and the nucleus in the IP "when they'd bought the dog" falls on the verb *bought*. These word(s) are already given and are considered old information. As a result, the nucleus falls on the previous lexical item or content word.

Wells (2006) also considers that the repetition of words does not necessarily constitute old or given information. We can also repeat old information using synonyms, in which we convey the same idea with various terms. These synonyms will also be deaccented, as exemplified below.

(8) Shall we wash the clothes? - Oh, I **hate** doing the laundry.
(9) Shall we walk there? - Yes, I **like** going on foot.

In the examples above (Wells 2006: 111) we can see that *doing the laundry* and *going on foot* are synonyms of *wash the clothes* and *walk* respectively, and, therefore, are deaccented.

The same applies to hypernyms of words or phrases already mentioned, which count as given information and make the nucleus go elsewhere (ibid.).

Taylor (2009: 119) is aware of this prosodic phenomenon and believes that emphatic prominence is used for augmentation purposes. However, this also extends to affective purposes, such as when the speaker knows the listener will comprehend but still employs extra emphasis. This may occur, for instance, as considered by this scholar, when the speaker wishes to express frustration or imply that the listener is stupid and, as a result, everything must be spoken slowly or clearly.

## 6. Objectives and Methodology

This research aims to find out whether four well-known online TTS software apply the anaphora rule and the deaccenting of given information to a corpus of 10 lines used as input. Each software will produce 10 audio files —40 in total— that will be analysed using *Praat*[2], a speech analysis software, with which we will be able to see the pitch contour of each utterance and determine whether the anaphora rule has been achieved or not. With this information, a qualitative and quantitative analysis will be performed to draw a conclusion on the performance of these software.

This section presents a comprehensive analysis of text-to-speech (TTS) software, examining its intricate operational aspects. Following that, the corpus under investigation will be presented. Furthermore, we will provide the audio speech software used to analyse the outputs produced by AI voices, along

[2] *Praat* it a speech analysis software developed by Paul Boersma and David Weenink from the University of Amsterdam. The version used in this analysis is 6.2.22

with a description of our methodology in implementing this programme to the audio files.

### *6.1 TTS software selected*

We have selected four online TTS software (Lovo, Murf, PlayHT, and Replica Studios), in which we have chosen two different synthetic voices to utter the sentences. The reason why we have chosen these software is because they are the top four TTS service providers on Google with free testing demos, which allows us to test their speech services. The voices for each software are one male with a British English accent and one female with American English. This way we can observe if there is any variation concerning regional accent.

One thing we have not been able to discover is the exact type of text-to-speech model, as seen in section 2, these providers offer. This information is not provided on the websites of these companies and, unfortunately, we cannot include it in this work. Nonetheless, given the development level of these software, we believe they are a combination of the models explained by Taylor (2009), and modern prosodic ones, like the CRF P2P model, combining linguistic, articulatory, and prosodic features.

#### *6.1.1 Lovo*

Lovo Inc. is a Delaware-incorporated company. According to the information provided on their website (www.lovo.ai), the software offers next-generation AI voiceover and a text-to-speech platform with human-like voices. It offers over 180 voice skins in 33 languages, each with unique traits for bespoke content. In addition, new voices are added each month. They offer advanced text-to-speech technology with "authentic voices. Truly human emotions in every voice created, breathing life into your content". In addition, their "mind-blowing voice cloning technology" requires just 15 minutes from users to create their customized voice skin.

Lovo provides several AI voice services: voice lab, a software to create dialogues with multiple characters; Lovo API, a service for companies in need of voices for their API (Application Programming Interface), like companies whose call centers need automation and create an IVR (Interactive Voice Response) to deal with customers' requests, of video game developers who need voices for game characters; and Lovo Studio, an online platform to create synthetic speech. From the three options available, we have used Lovo Studio to analyse the deaccenting of given information. The voices selected to deliver the lines under study are Shawn Prince, a male senior voice actor with a British English accent; and Cameron Sorenson, a female senior voice actor with an American English accent.

### *6.1.2 Murf*

Murf.ai is a Utah-based company whose AI-enabled text-to-speech online application enables users to make "human-like" voiceovers for movies and presentations without the need for a voice actor or complicated recording equipment. According to their website, Murf is attempting to streamline speech audio and make high-quality voiceovers accessible to all users. Its motto as we open the website states: "AI-enabled, real people's voices".

Murf offers different services including text-to-speech, voice cloning, voice-over video, voiceover Google Sides add-on, voice changer, and API. For this work, we have used the text-to-speech service and the online platform is similar to Lovo's.

Unlike Lovo, Murf allows us to modify the pitch and the speed of the output delivery, and also insert a pause (extra weak, 250ms; weak, 0.5s; strong, 1s; extra strong, 1.25s). To insert a pause, we just need to place the cursor where we want to include it, click on *add pause*, and select the type of pause we wish to add. Finally, we click on the *play* button, and the audio file is automatically generated.

An interesting feature this software offers is that when we place the cursor over the play button, next to it, a dialogue box

icon with a plus sign appears. If we place the cursor over this icon, the word "Emphasis" pops up. When we click on it, a new window opens displaying the line we are editing on a two-axis graph. The vertical axis ranges from "high" at the top, "0" in the middle, and "low" at the bottom. Each word is isolated on the horizontal axis, and each word has a node or slider that can be moved up or down depending on the emphasis we want to give to each word. Once the word or words that need to be highlighted are set, we can preview the audio to check whether we like the result or not, and, eventually, apply or cancel the changes. This option is really interesting and we will use it to de-emphasise the words that should be deaccented in our sentences to compare the result with the standard reading provided by this very software.

Finally, this software offers a "voice changer" option, with which we can upload our own recorded audio files. Then, the software transcribes the audio and assigns a voice skin to voice it for us. This text and voice skin can be edited afterward on the workstation just any other text and voice.

The voices selected from Murf to deliver the lines under study in our project are Anna, a young adult female voice with an American English accent; and Edward, a middle-aged male voice actor with a British English accent.

### *6.1.3 PlayHT*

PlayHT is an online company, based in Bangalore (India), which offers "high-quality text-to-speech synthesis and audio accessibility solutions using the most realistic AI voices in the world", according to their website. We can choose from 142 languages and accents. In addition, this company also offers a voice cloning option, similar to the one we saw in Murf. PlayHT displays examples of voice clones with realistic synthetic voices of well-known celebrities such as Tom Hanks, John F. Kennedy, and Elon Musk, among others. The TTS software developed by PlayHT is named "Peregrine", a "TTS method which is able to synthesize speech with a higher degree of realism, making it ba-

sically undistinguishable from natural speech as spoken by humans", as they declare on their website.

Once we type in the line we want to voice, the software generates the audio file. Moreover, if we wish to have different versions of the line, there is an option called "Re-Generate Previews", with which we can generate new audio files with different deliveries. Then, we select our preferred version, and we can download the WAVE audio file on our computer.

From the voices available on PlayHT, we have selected Arthur, a male adult voice actor with a British English accent, and Evelyn, a female adult voice with an American English accent.

### *6.1.4 Replica Studios*

Replica is a Techstars company based in Dover (USA) and Brisbane (Australia). Replica's website (replicastudios.com) offers "AI voice actors for games, film, and the metaverse" as stated on the website homepage. With Replica, a skilled voice actor spends countless hours teaching their AI how to perform. Then the AI learns how to perform by imitating the unique speech patterns, pronunciation, and emotional range of human voice actors. The final product is an AI voice actor that can be used in games and films.

Unlike Lovo and PlayHT, Replica does not provide voice-cloning services, only synthetic voice acting. Moreover, it only offers voices in English with multiple regional accents. Replica's unique feature is that to work with its software we must download their bespoke app, which works both for Windows and Mac computers. Just like in the three previous software, we have to create an account with our email address to start working with it.

When a new project is created, we need to select one AI voice actor to play our character, for whom a name and a biography can be given. Then a list of 109 AI voice actors is provided from which we can select the perfect voice for our character. Once we feed the text, we can go to the "Character" tab for that line and select from a drop-down menu the character (previously creat-

ed) that will voice the line. Also, if we go to the "Style" menu, we can notice it will say by default "light-hearted", but there is a drop-down menu from which we can select two more styles: angry, and sad. This is quite interesting since we can change the style of the delivery without retyping the line. Finally, we can download the audio file in different formats

This software offers a wide range of voices with over 100 AI voice actors from which we can select the most suitable ones for our projects. In our case, we have selected Atlas, a senior (+55) male voice with a British accent and light-hearted style; and Amber, a young adult (18-34) female voice with an American English accent to voice our lines.

### *6.2 Corpus*

The input corpus used in this study comprises the ten sentences listed in Table 2. The given information that should be deaccented, as predicted by the anaphora rule described in section 5, is underlined in each example, and the nucleus, where the tonic syllable should fall, is highlighted in bold.

| Number | Line |
|---|---|
| Line 1 | Do you want a room with a bath or wi**thout** a bath? |
| Line 2 | You won't tell him, **I** won't tell him. |
| Line 3 | In this world perhaps, but in **my** world the books will be nothing but pictures. |
| Line 4 | I know how she feels, but how do **you** feel? |
| Line 5 | The bread isn't yesterday's, it's the day be**fore** yesterday's. |
| Line 6 | Liverpool three, **Ar**senal three. |
| Line 7 | A banana that has dots in it is tastier than a **green** banana. |
| Line 8 | In a sense he would have hoped Smeaton hadn't confessed, but Smeaton **did** confess. |

| | |
|---|---|
| Line 9 | It's a question of overall numbers and at the last elections the conservatives said they would cut immigration from hundreds and thousands a year to **tens** of thousands a year on a net basis. |
| Line 10 | No woman had ever made that step from royal mistress to the throne, getting the Queen, a **real** queen, out of the way. |

Table 2. List of lines used as input in the TTS software

### *6.3 Speech audio software*

The software used in this paper to check the intonational accuracy of the TTS described above is *Praat*, a software that allows us to display the audio wave, and the pitch contour of a given utterance and also to write the words for the utterances placing them accordingly on the horizontal axis (time) on screen. Thus, we can get a clear picture of the whole utterance, which can eventually be exported in different formats, such as pdf, png, etc. Image 1 illustrates the kind of representation produced by Praat, which will be the basis of our analysis.

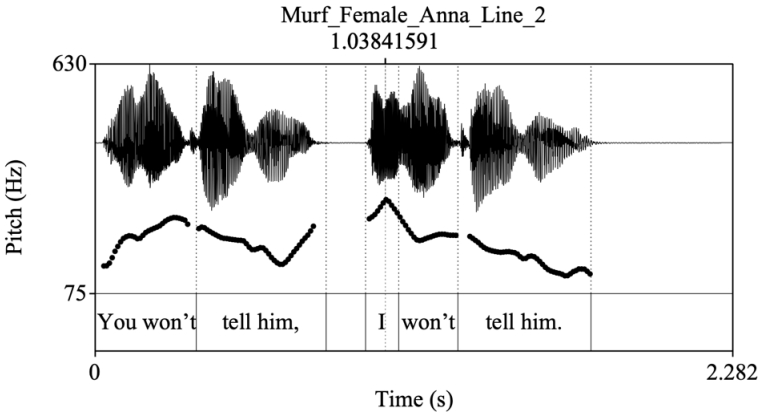



Image 1. Pitch contour and audio wave for Line 2 uttered by Anna from Murf

As shown in the image, Pitch (measured in Hz) is displayed on the vertical axis, while Time (in seconds) appears on the horizontal axis. The audio wave for the utterance is shown on the top of the image, its corresponding pitch contour or melody appears just below, and at the bottom, we can see the words of the utterance, which coincide with the melody and the audio wave.

This example belongs to line 2 (*You won't tell him, I won't tell him*), uttered by the female voice Anna from Murf, and shows a satisfactory case of deaccentuation of given information. In the second IP "I won't tell him", we can hear and see (following the pitch contour) that "I" is the nucleus of the IP, and "won't tell him" is the tail. Thus, "won't tell him" has been deaccented, since it has been uttered in the first IP and is, considered old information. Consequently, "I" is accented following the narrow focus structure.

If we now examine the example shown in Image 2 below, we can identify an unsuccessful case of deaccenting. Here, we have the same utterance (line 2) as in Image 1 voiced by Shawn, from Lovo. This time, the voice skin emphasized the word "won't", instead of "you", placing the nucleus in the wrong word. This, undoubtedly, alters the illocutionary force and the pragmatic meaning of the utterance. As can be seen in Image 2, the melodic contour is flat on "you" and suddenly soars over "won't", after which it declines to provide the tail ("tell him").

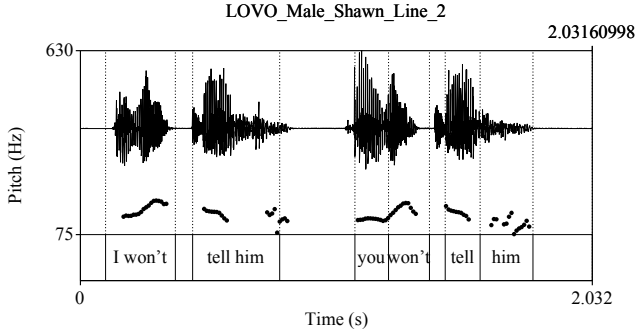



Image 2. Pitch contour and audio wave for Line 2 uttered by Shawn from Lovo

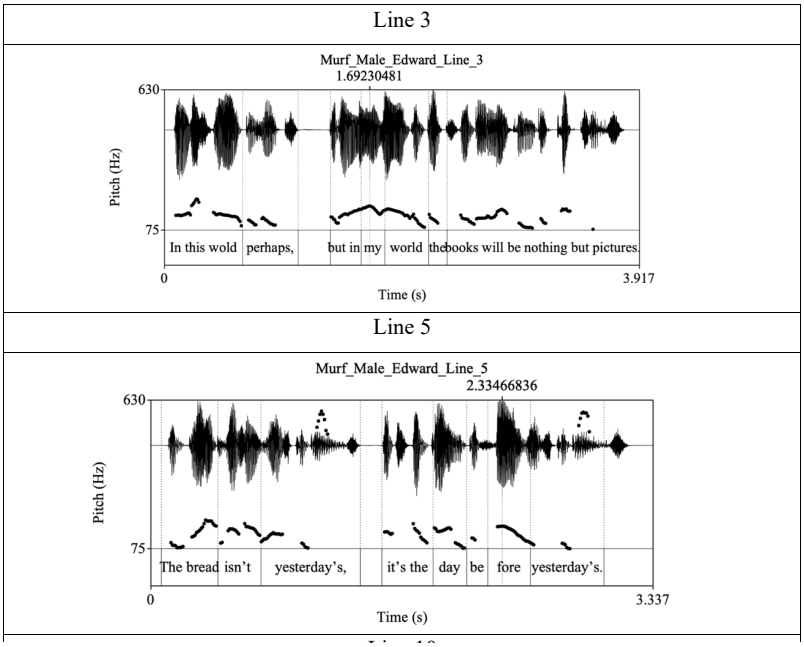
## 7. Analysis and Results

In this section, we will analyse the intonational representation of the utterances produced by the synthetic voices for the sentences in Table 2 where the anaphora rule applies. Those utterances that are not subject to this will not be shown in full since they are not the focus of this study. Nonetheless, all the contours can be accessed upon request. Finally, we will conduct a quantitative analysis and get the overall numbers for each software.

### *7.1 Utterances that successfully apply the anaphora rule*

#### *7.1.2 Murf*

The male synthetic speaker called Edward applies the anaphora rule to utterances 3, 5, and 10, as shown in Table 3.



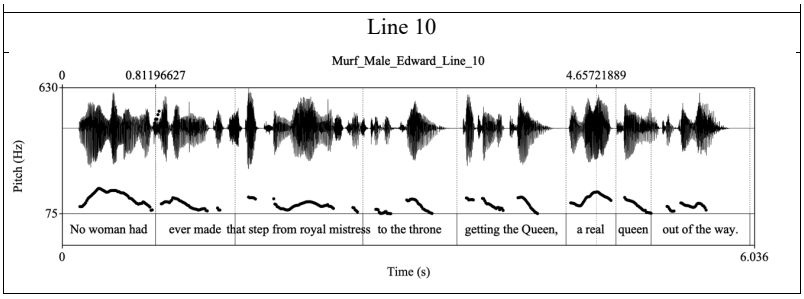
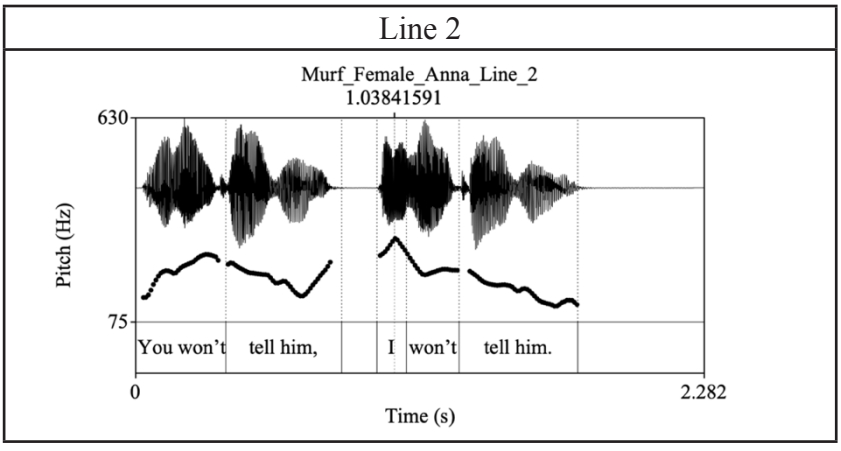



Table 3. Lines successfully uttered by Murf's male voice Edward.

As can be seen, in line 3, for the IP "but in my world" the nucleus falls on "my", thus deaccenting "world", which is old information as it comes from the previous IP. Next, in line 5, for the IP "it's the day before yesterday's" the nucleus falls on the second syllable of "before" and deaccents "yesterday's". And the same applies in line 10 for the IP "a real queen", in which "real" gets the tonic syllable and "queen" is deaccented anaphorically.

Murf's female synthetic speaker Anna applies the anaphora rule in lines 2, 7, and 10, as shown below.



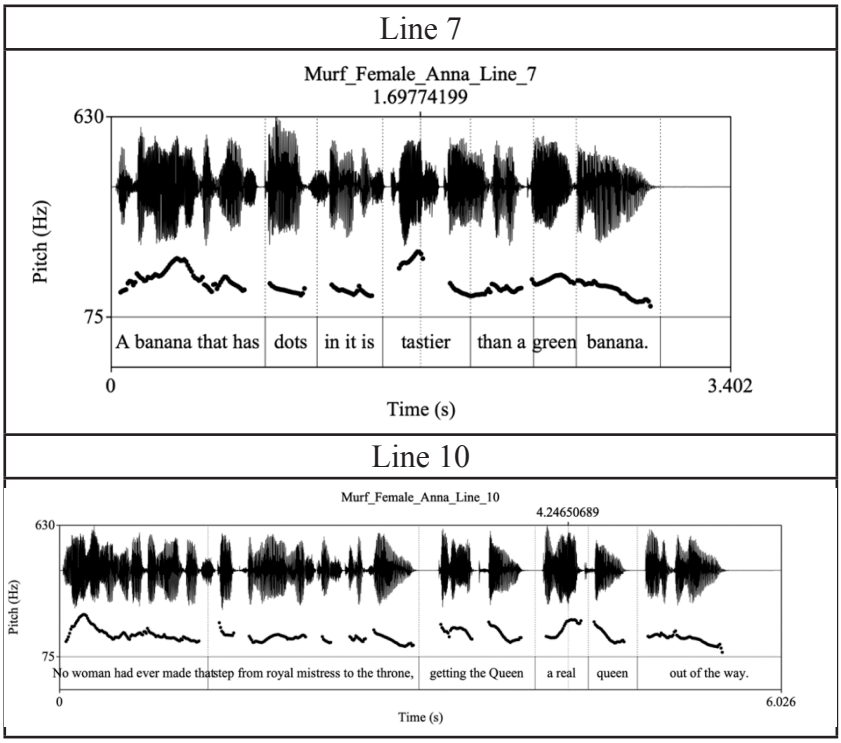



Table 4. Lines successfully uttered by Murf's female voice Anna.

In line 2 we can see how "I" gets the nuclear accent and is emphatically accented, while "won't tell him" is deaccented as old information and the pitch line decreases gradually. In line 7, for the IP "than a green banana", "green" gets the tonic syllable, and "banana" is deaccented. Finally, in line 10, for the IP " a real queen", "queen" is deaccented as old information, and "real" gets the tonic syllable in narrow focus.

### *7.1.3 Lovo*

The delivery of the lines by the voices selected from Lovo is as follows: the male synthetic voice Shawn successfully uttered three lines (4, 6, and 7). However, the female voice skin Cameron did not apply the anaphora rule to any of the lines.

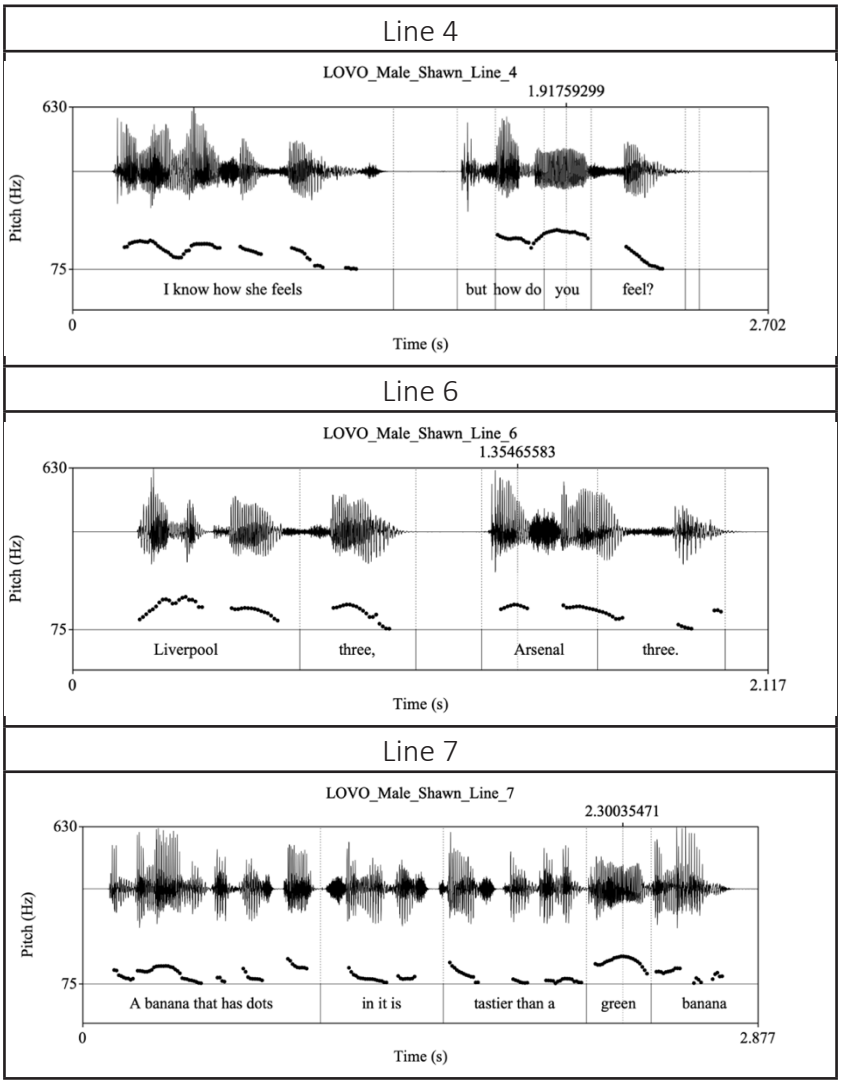



Table 5. Lines successfully uttered by Lovo's male voice Shawn

In line 4, the tonic syllable falls on "you", and "feel" is deaccented as given information. The same applies to line 6, in which "three" is deaccented in the IP "Arsenal three", although it received the tonic syllable in the previous IP "Liverpool three".

Being given information, "three" is deaccented in favour of "Arsenal", which gets the nucleus on its first syllable. In line 7, in the IP "than a green banana", "green" receives the tonic syllable, and "banana" is, therefore, deaccented as this item of information is known.

### *7.1.4 PlayHT*

The voices selected from PlayHT delivered the lines in the following fashion: the male voice (Arthur) successfully uttered two lines (3 and 9), while the female voice (Evelyn) did not apply the anaphora rule to any of her lines. Arthur's lines are shown in Table 6 below.

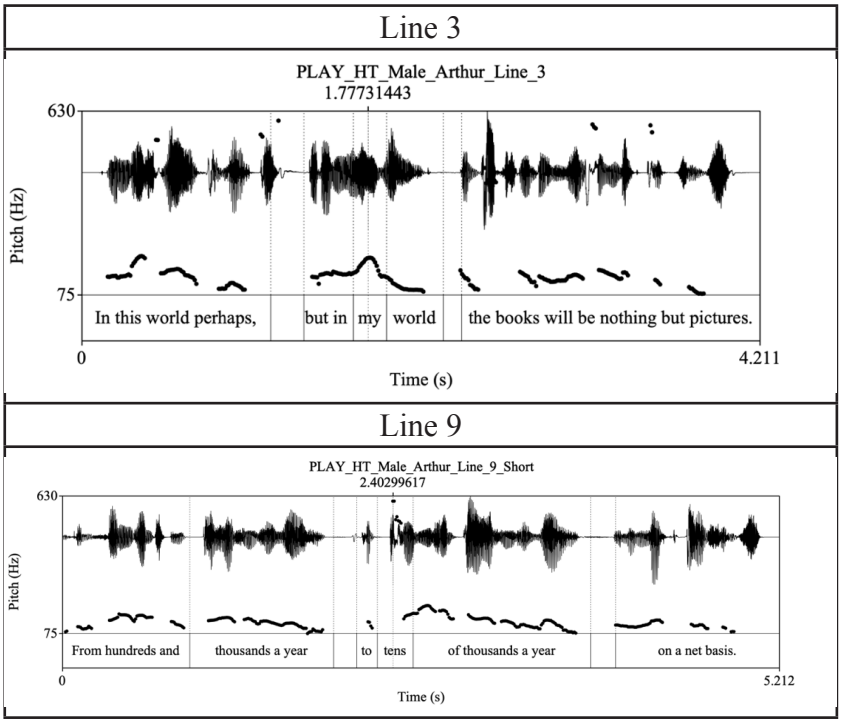
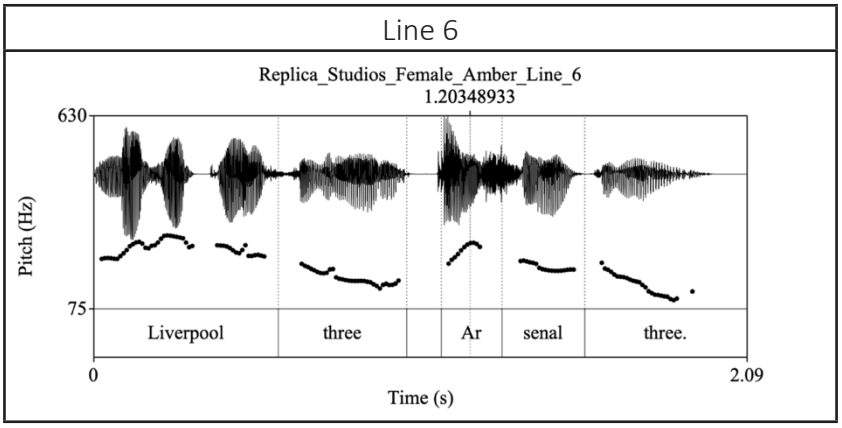



Table 6. Lines successfully uttered by PlayHT's male voice Arthur.

We can see that in line 3, for the IP "but in my world", the word "my" is accented and marked emphatically, thus deaccenting "world", which is old information. This way, the contrast is clear and the narrow focus is achieved by the anaphora rule. The same happens in line 9, in which "tens" is the tonic syllable and is highlighted to the detriment of "of thousands a year" which is treated as known information and, consequently, is deaccented. The pitch contour for "tens" goes as high as 630 Hz and could be difficult to see since it is somewhat blurred by the audio wave. Nonetheless, we can see the dotted mark at second 2.4 in the audio graph.

### *7.1.5 Replica Studios*

The delivery shown by the synthetic voices from Replica Studios is as follows: the male voice (Atlas) did not succeed in applying the anaphora rule to any of his lines, while the female voice (Amber) achieved the deaccenting of given information in 2 lines (6 and 10). Amber's lines are displayed in Table 7 below.



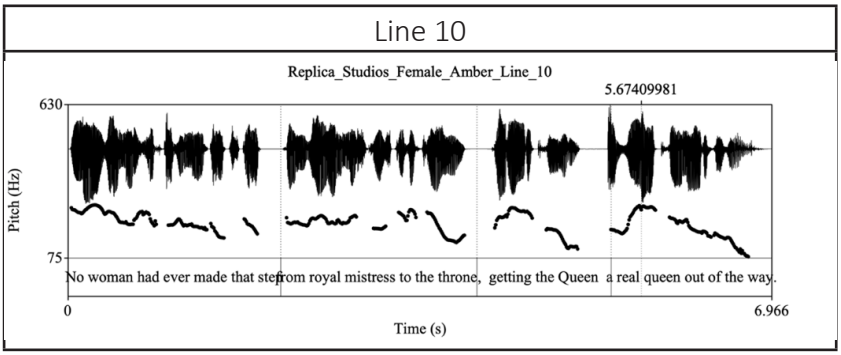



Table 7. Lines successfully uttered by Replica Studio's female voice Amber.

In line 6, it is clearly seen that the pitch contour for the first syllable of "Arsenal" is where the nuclear accent falls, and from there to the end of the utterance, the pitch contour decreases gradually along the tail of the IP. Thus, "three", which is information given in the previous IP, is deaccented and this, as explained by Mateo (2014: 120), allows the hearer to predict information. Finally, line 10 shows narrow focus since "real" receives the tonic syllable and "queen", which comes from the previous IP and is considered known information, is deaccented, and starts the tail of the IP. Here, again, the pitch contour clearly illustrates this phenomenon, with a pitch soaring from "a" to "real" and then a gradual decrease until "way", where the tune ends.

### *7.2 Quantitative results*

After discussing the qualitative results of the delivery of the utterances by the different voice skins selected from the TTS software, we will now turn to how that information can be translated into overall numbers to have a better picture of the situation concerning this prosodic phenomenon.

The most successful TTS software in terms of the anaphora rule is Murf, which achieves deaccenting of given information in 33.3% of the utterances (6 out of 20 lines). Next is Lovo with 15%

(3 lines out of 20), and finally, PlayHT and Replica Studios both share a 10% success rate (2 lines out of 20). This information is shown in the graph below.

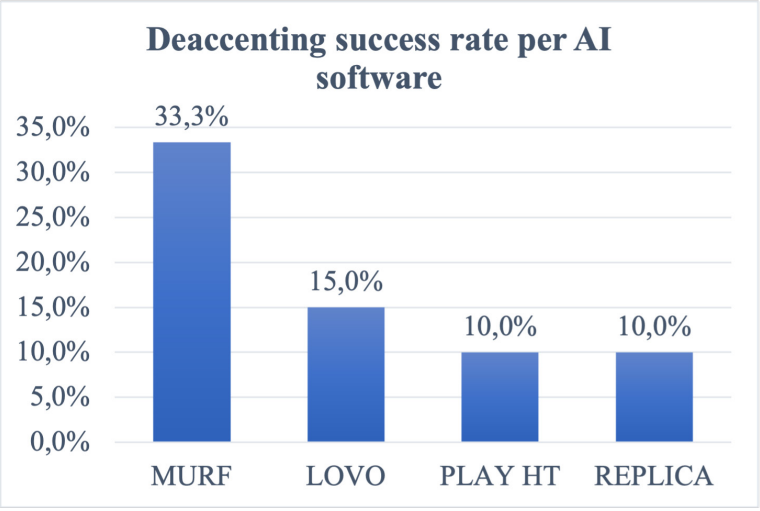



Figure 2. Deaccenting success rate per AI software

In terms of accent, all male speakers had British accent, and all female voices had American accent, so the numbers that will come next will be valid both in terms of gender and accent. The information retrieved shows that the deaccenting success was higher for the male British speakers than for the female American speakers in general numbers as shown in the graphs below.

The information from the graphs below shows that the deaccenting success rate, although low, is higher for the male speakers with British accent (20%) than for the female Americans (12.5%). In terms of TTS software, we can see that in Murf both the male and female speakers reach a success rate of 30%. In Lovo, only the male speaker achieves deaccenting (30%), while the female voice does not. And the same applies to PlayHT, where only the male voice applied the anaphora rule (20%). The opposite takes

place with the voices from Replica Studio, as only the female voice succeeds in applying the deaccenting of given information 12.5% of the time.

Finally, the overall numbers for the whole TTS software and speakers show that the anaphora rule was only applied 32% of the times, which means that 68% of the utterances did not have the illocutionary force, or pragmatic load, successfully delivered, as shown in the pie chart below.

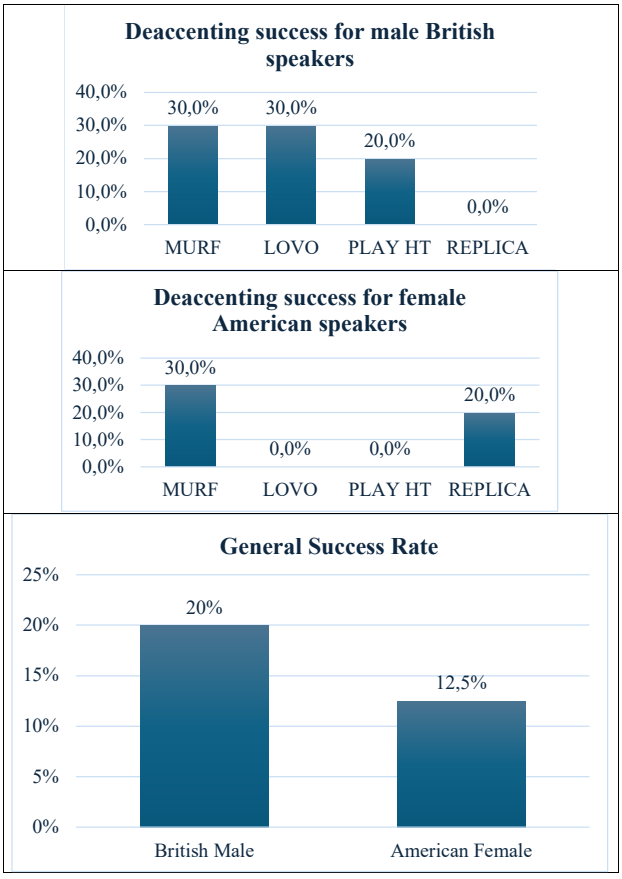



Table 8. Success rate according to gender and accent


Figure 3. Anaphora rule success for all the TTS analysed.

***7.3 Murf's emphasis tool.***

We have also analysed how Murph's emphasis tool works and whether or not it adds the desired emphasis to the word we want to highlight. We have tested this feature with two of the lines in which the female voice Anna did not apply the anaphora rule (lines 4 and 5). We have manually highlighted the words to emphasise and created new audio files for the two voices (Anna and Edward). Table 9 below shows how the software allows us to manipulate the pitch contour for each element in the utterance.


Table 9. Murf's Emphasis tool.

The red lines in the images shown in Table 9 belong to the pitch contour of the utterance in the software. We can click on it at whichever point we want and drag it upwards to raise the pitch for that specific part of the utterance. This way, we manually order the software to apply emphasis to that specific element. Once we have applied the emphasis, we download the audio files and proceed to analyse them in Praat.

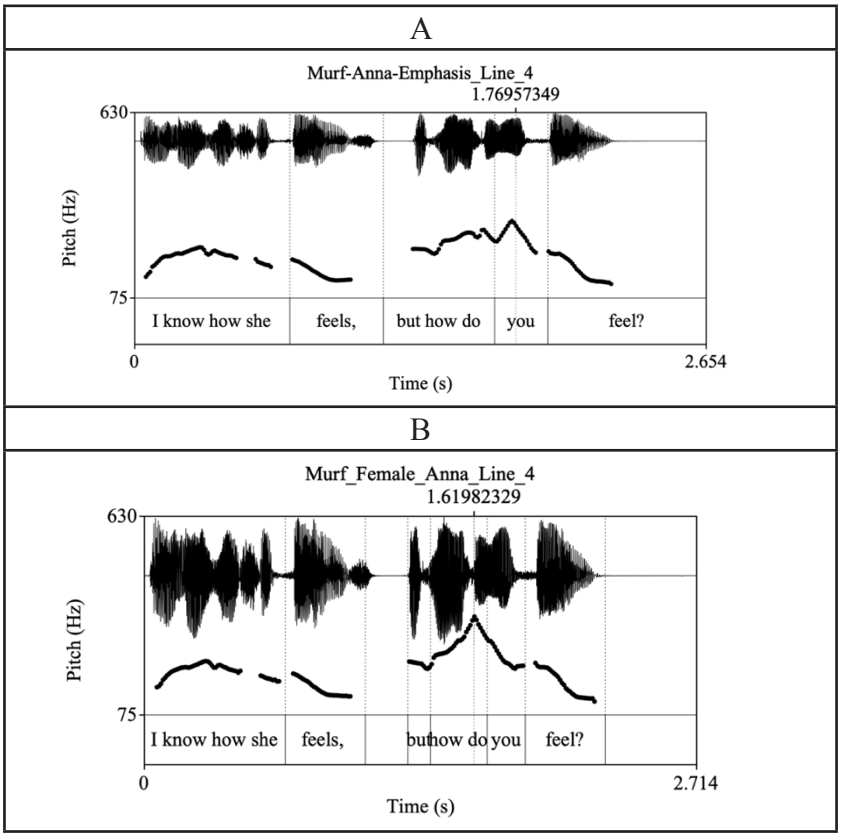



Table 10. Pitch contour with and without emphasis for line 4 voiced by Anna

Table 10 shows the pitch contour for line 4 voiced by Anna. Image A shows the utterance with manipulated emphasis, and as can be seen, the pitch contour rises on the word "you", thus producing the desired effect that the anaphora rule produces. Image B, at the bottom of the table, shows the utterance delivered without emphasis, and the way the software produced it right after typing in the line. The pitch contour shows that the highlighted word is "do", which means that the anaphora rule is not applied, and the pragmatic load is not successfully delivered.

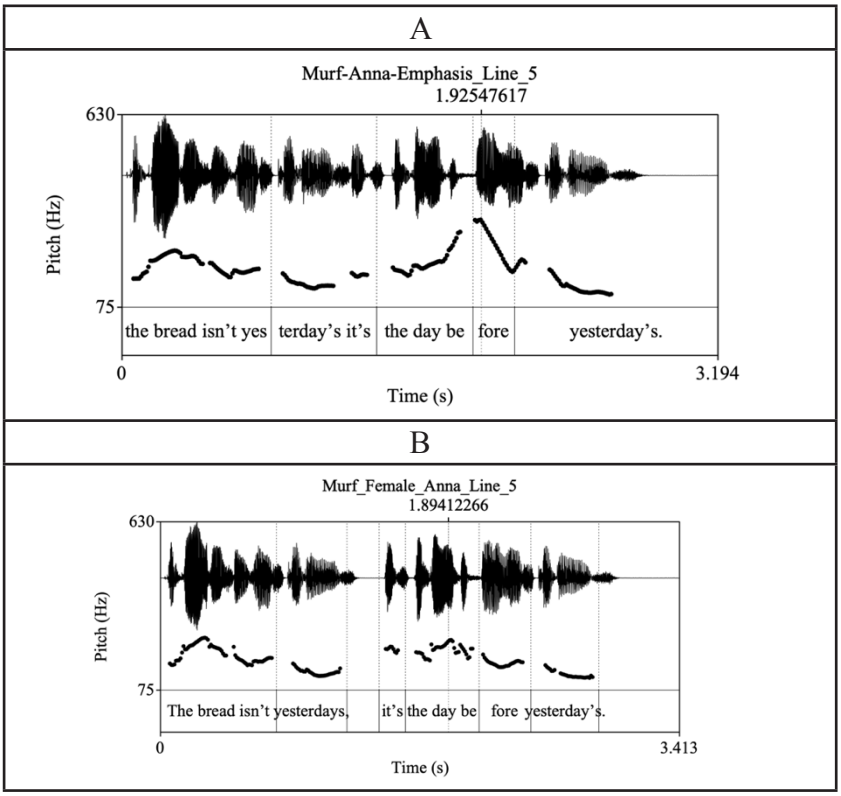



Table 11. Pitch contour with and without emphasis for line 5 voiced by Anna

Table 11 shows Anna's delivery of line 5. Image A shows the delivery of the utterance after consciously manipulating the emphasis with the software's toolkit. As shown, the pitch contour rises over the second syllable of "before" and makes it the tonic syllable, thus deaccenting "yesterday's" as expected by an English native speaker. However, in image B, which shows the pitch contour for the utterance as it is entered into the software, the tonic syllable falls on "day". That is unexpected since it applies narrow focus on a word that should not be highlighted rather than broad focus. It seems as if the software knew it should not

highlight "yesterday" in broad focus and tried to apply narrow focus, making a mistake by emphasizing the wrong word.

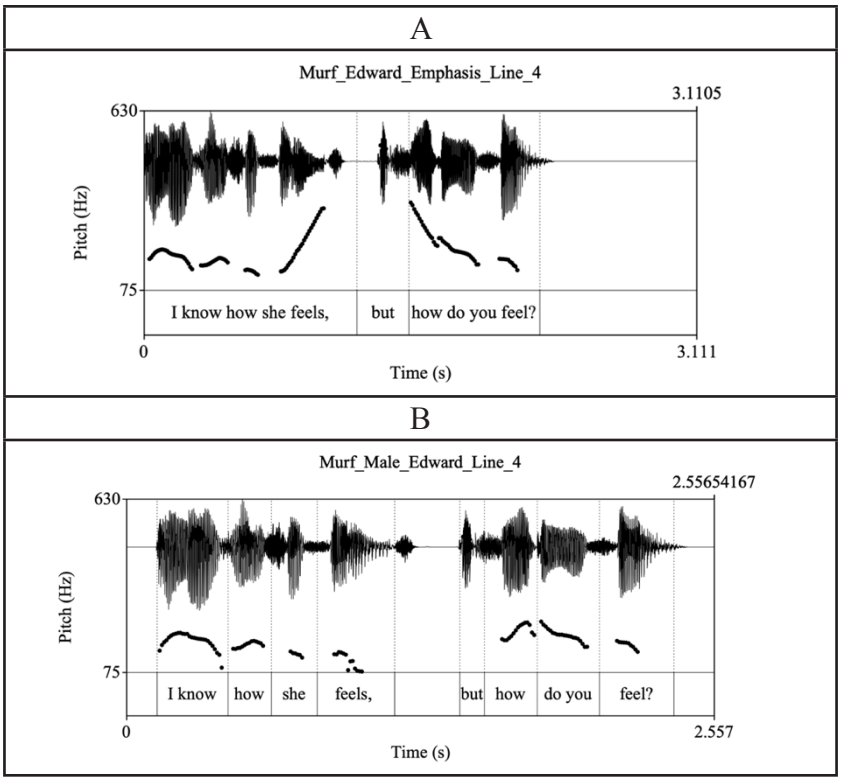



Table 12. Pitch contour with and without emphasis for line 4 voiced by Edward

If we now compare the utterances produced by the male voice skin Edward for line 4, shown in Table 12, we can see that the result is opposite to that of Anna's. After manually introducing the emphasis on the word "you", as we did with Anna, the software produced an utterance with a different emphasis, as shown in image A. The emphasis falls on "how" instead of "you" and, therefore, does not produce the desired effect of the anaphora

rule. Image B shows the delivery of the utterance without emphasis, and the contour shows the same pattern as in A. The tonic syllable falls on "how" and there is no anaphora rule.

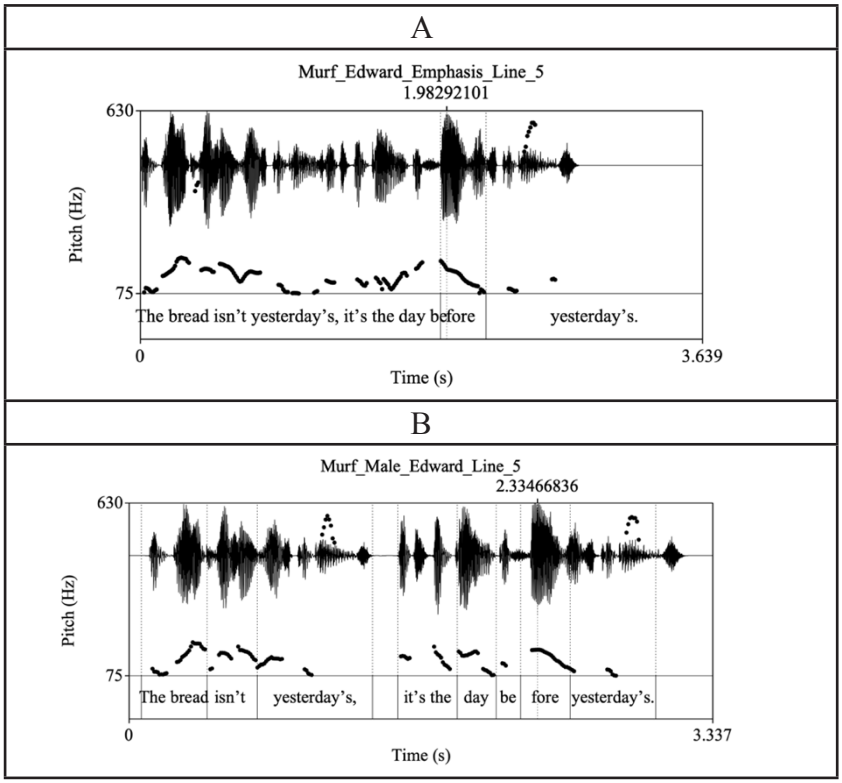



Table 13. Pitch contour with and without emphasis for line 5 voiced by Edward

Line 5 shows a successful result after manually manipulating the emphasis for Edward, as shown in image A. This time, the software had no trouble highlighting the selected word and the result is satisfactory as it applies the anaphora rule. Image B shows the delivery of the utterance without manipulating the emphasis with the software's toolkit, and —as already comment-

ed in section 7.1.2—, Edward was able to produce the anaphora rule from scratch.

**8. Conclusion**

With the results obtained from the analysis, we can conclude that these TTS software do not successfully apply the deaccenting of given information which is unconsciously applied by English native speakers to highlight the relevant information in spoken conversation. This means that although the level of catenation and voice quality of the voice skins are acceptable and can sound natural and believable to people's ears from a segmental perspective, the prosodic features that add pragmatic value to the utterances, like the anaphora rule, and which belong to augmentative prosody, have not been mastered yet. The TTS companies we have studied offer services with catchy slogans like *lifelike voices*; *natural-sounding voices*; *truly human emotions*; and *perfect voiceovers*. Nonetheless, we have seen that a prosodic feature such as the anaphora rule is applied only 32% of the time, and that is the opposite of showing truly human emotions. In fact, that may sound unnatural and non-human.

Failing to apply the deaccenting of given information could be problematic in cases in which TTS voices deliver audio for e-learning modules, audio guides, and audiobooks, for instance, as they do not focus on the right information when they deliver the audio for the text. In those cases, the hearers would not benefit from the service being provided, since the right information is not highlighted, and the client paying for the TTS service would have his message not delivered in full. Here is where the debate between human voices and AI synthetic voices still favours humans. However, we have seen that TTS companies like Murf allow the user the chance to highlight specific words in utterances to apply a narrow focus. This tool seems to work well and applies the emphasis on the desired words, although sometimes it does not, as shown. This tool seems useful for short phrases, but

we believe that it could be complicated to use it in long texts like audio books or instruction manuals.

This study has been limited to four companies that deliver text-to-speech services and two speakers for each company. More research should be conducted studying other companies and voices to find whether the figures shown in our work apply to those or not. Moreover, other prosodic features used by synthetic voices, like tone direction, tonality, pitch, etc., and their application to TTS should be studied, following the research of Rodríguez Fernández-Peña (2023a, 2024), to find whether machines are finally getting human or not.

Voice-over agencies with human voice talents in their databases already provide synthetic TTS services, like voicebooking.com and voicearchive.com, which seems an indicator of the direction the professional voice-over and dubbing industry is taking, as explained by Rodríguez Fernández-Peña (2023b). In addition, thanks to deepfakes, actors and other celebrities can reach an agreement with AI companies to perform when they are unavailable or even after they have passed away using their voices and physical appearances. The presence of AI voices in everyday life is already a reality, it is here to stay, and big brands are spending important sums of money on it. Nonetheless, prosody seems to be a unique human-like feature that machines cannot execute yet.

## 9. References

Aguero, P. D., & Antonio, B. C. (2003). "Phrase break prediction: a comparative study". *Procesamiento del lenguaje natural, 31*, 107-114.

Ashby, P. (2011). *Understanding Phonetics* (Vol. Understanding Language Series). (B. Comrie, & G. Corbett, Eds.) London: Hodder Education.

Austin, J. L. (1962). *How to do things with words.* Oxford: Oxford University Press.

Cohen, M., Giangola, J.P., Balogh, J. (2004). *Voice User Interface Design*. Addison-Wesley Professional.

Collins, B., & Mees, I. M. (2013). *Practical Phonetics and Phonology. A resource book for students.* Oxon, United Kingdom: Routledge.

Cruttenden, A. (2014). *Gimson's Pronunciation of English* (Vol. 8). Oxon (UK) and New York (USA): Routledge.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Mons, Belgium: Springer Science+Business Media Dordrecht.

Dutoit, T. (1997b). *High-quality text-to-speech synthesis: an overview*. Journal of Electrical and Electronics Engineering Australia, 17(1), 26-36.

Estebas-Vilaplana, E. (2014). "Phonological models of intonational description of English". In R. Monroy-Casa, & I. Arboleda-Girao, *Readings in phonetics and phonology* (pp. 231-260). Valencia: IULMA (Institut Universitary de Llengües Modernes Aplicades) Universitat de Valencia.

Halliday, M. (1967). "Notes on transitivity and theme in English, Part 2". *Journal of Linguistics (3)*, 199-244.

Hatim, B., & Mason, I. (1990). *Discourse and the translator*. Longman.

Hirschberg, J. (2006). "Pragmatics and Intonation". In L. R. Horn, & G. Ward, *The Handbook of Pragmatics*. Oxford: Blackwell Publishing.

Ilona, K., Gábor, O., & Péter, O. (2000). Prosody Prediction from Text in Hungarian and its Realization in TTS Conversion. *International Journal of Speech Technology, 3*, 187-200.

Klatt, D. H. (1987). "Review of text-to-speech conversion for English". *The Journal of the Acoustical Society of America, 82(3)*.

Mateo, M. (2014). "Exploring pragmatics and phonetics for successful translation". *(VIAL) Vigo International Journal of Applied Linguistics (11)*, 111-135.

Mott, B. (2011). *English phonetics and phonology for Spanish speakers.* Barcelona: Publicacions i Edicions de la Universitat de Barcelona.

Prince, E. F. (1981). "Toward a taxonomy of given/new information". *Radical Pragmatics*, 223-255.

Qader, R. L. (2017). "Statistical Pronunciation Adaptation for Spontaneous Speech Synthesis". *Text, Speech, and Dialogue: 20th International Conference, TSD* 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20 (pp. 92-101). Springer International Publishing.

Rodríguez Fernández-Peña, A. C. (2023a). "AI is great, isn't it? Tone direction and illocutionary force delivery of tag questions in Amazon's AI NTTS Polly". *Estudios de Fonética Experimental, 32*, 227–242. https://doi.org/10.1344/efe-2023-32-227-242

Rodríguez Fernández-Peña, A. C. (2023b). "Online cloud dubbing: How home recording stormed the dubbing industry". *Revista Tradumàtica. Tecnologies de la Traducció, 21,* 028-048. https://doi.org/10.5565/rev/tradumatica.335.

Rodríguez Fernández-Peña, A. C. (2024). "La desacentuación anafórica en inglés de las voces sintéticas de inteligencia artificial de Amazon Polly: un estudio de caso". En *IA, educación y medios de comunicación: modelo TRIC*. Dykinson SL.

Tahon, M. L. (2017). "Perception of expressivity in TTS: linguistics, phonetics or prosody?" *Statistical Language and Speech Processing. 5th International Conference on Statistical Language and Speech Processing,* October 23-25, 2017 (pp. 262-274). Le Mans, France: Springer.

Tan Xu, C. J.-Y. (2022). Natural Speech: End-to-End Text-to-Speech Synthesis with Human-Level Quality. arXiv.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

Tench, P. (2009). "The Pronunciation of Grammar". *3rd International Congress on English Grammar*. Salem, TN, India: Sona College of Technology.

Wells, J. (2006). *English Intonation: an introduction.* Cambridge: Cambridge University Press.

Yule, G. (1996). *Pragmatics*. Oxford: Oxford University Press.