POETRY AND SPEECH SYNTHESIS: SPARSAR RECITES

Rodolfo DELMONTE

ABSTRACT • In this paper we present SPARSAR, a system for English poetry recital. The system is parasitic on TextToSpeech (TTS) systems available both online and on Macintosh computers. It creates prosodic parameters and phonetic transcriptions on any input text to be used by the TTS in order to normalize and improve current systems which are statistically based. In order to show TTS inability to produce semantically coherent and expressive readings Italian texts will be used at first and critical points indicated and discussed. Then SPARSAR architecture will be introduced and its three layers presented in detail. The ability of the system to generate appropriate prosodic parameters will be discussed in relation to a poem by Sylvia Plath, Edge. The peculiarity of this poem is its richness in enjambments, which are not captured at all by statistically based TTS. Eventually, latest work on Elizabethan poetry will be presented and a Sonnet by Shakespeare will be transcribed and annotated with prosodic parameters' values by the system; in particular, it will be shown how the lack of a specific component to account for contractions and rhyming violations makes best commercial TTS systems even unable to pronounce words correctly.

KEYWORDS • TextToSpeech; Prosody; NLP; Poetry

1. Introduction

TTS or a TextToSpeech system has been around for quite a number of years now, from about the end of 60's. However, only in the last four or five years it has become a companion to most mobile phones and computer systems and applications. Thanks to its easiness of usage and implementation, digital voice assistants are more and more approaching a human-like appearance thanks to presence of TTS applied to virtual talking heads or agents. This notwithstanding, TTS has not yet reached a level of technological maturity comparable to that of Automatic Speech Recognition (hence ASR) which attained its apex when announcing Continuous Speech Recognition with Large Vocabularies already in the 90's. However, we feel that, as happened in ASR, a combination of phonetic and linguistic structural information with proper use of statistical procedure will eventually lead to improved results in TTS (see Delmonte, 2008).

Even though the introduction and practical usage of TTS in Voice Assistants and Dialogue Systems is an encouraging result, we feel that there are two issues that need to be considered: there has been an enormous improvement in the quality of speech thanks to the use of statistical approaches based on large training corpora, but the counterpart to that is the lack of linguistic knowledge to allow generalizing its usage. Pros and cons of this kind of attitude towards TTS may be summarized by using R. Sproat and J. van Santen opinions as expressed in the Introduction (Sproat 1997):

We feel it is nevertheless important to point out that the ultimate goal - that of accurately mimicking a human speaker - is as elusive as ever and that the reason for this is no secret. After all, for a system to sound natural, the system has to have real-world knowledge, know rules and exceptions of the language; appropriately convey emotive states; and accurately reproduce the acoustic correlates of intricate motor processes and turbulence phenomena involving the vocal cord, jaws and tongue. What is known about these processes and phenomena is extremely incomplete, and much may remain beyond the grasp of science for many years to come.

In view of this, the convergence in current work on TTS is perhaps somewhat disturbing. For example, a large percentage of current system use concatenative synthesis rather than parametric/articulatory synthesis. We believe that this is not for theoretical reasons but for practical reasons: the quality levels that are the current norm are easier to attain with a concatenative system than with a parametric/articulatory system.

However, we feel that the complex forms of coarticulation found in human speech ultimately can only be mimicked by accurate articulatory models, because concatenative systems would require too many units to achieve these - significantly higher - levels of quality.... We make these remarks to express our belief that TTS is not a mature technology. There may exist standard solutions to many TTS problems, but we view these solutions only as temporary, in need of constant scrutiny, and if necessary rejections. [ibid., 1-2]

Current successful TTS systems go through a choice of the most appropriate diphone unit or speech segment by means of statistical measurements (mostly HMMs) based on a database of speech segments collected and annotated in advance for that purpose (Delić et al. 2017). Here “appropriate” means not only containing the legal inventory of phonetic segments, or phones, of the language with all its “most relevant” co-articulatory phenomena; but also reflecting “most” prosodic features of the language/domain - “all” would be impossible due to the number of segments required in order to reach statistical significance (Black et al. 2011; Black et al. 2012).

The quantity of information needed to achieve optimal performances in TTS is simply too much to be realistically available due to the “sparseness” problem: too many variables have to be taken into account in order to produce a sensible mapping of all linguistically significant and perceptually relevant parameters to reach comparable statistical results in terms of naturalness of speech output. However, current corpus-based TTS systems try to approximate the best possible selection of acoustic units on the basis of greedy algorithms, an optimal selection of text to be used as reference corpus by the speakers and a model-based greedy selection of units. By means of such an approach prosodic information is preserved and reproduced to the best possible approximation, which however is not sufficient to guarantee a decent level of naturalness.

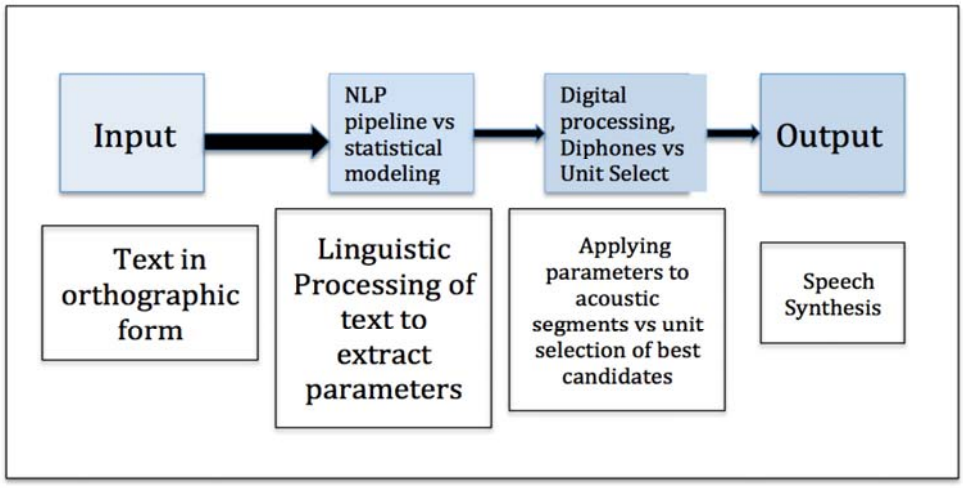


Figure 1: Pipeline of TTS system for speech synthesis

The traditional way to read a text by a speech synthesizer is represented in Figure 1 above. As can be clearly noted, parameters are created by the linguistic module which builds some kind of representation on top of the input text to be read aloud. Current statistical TTS are trying to do away with the intermediate linguistic processing module and use a different approach which is described in (Rajeswari 2012; Archana 2013). They start from a speech-signal corpus database which is used in the training phase by means of HMM-based machine learning procedures. The output of this procedure is meant to be a speech waveform with all needed linguistic parameters. The approach is very similar to the one used for speech recognition except for the final step which is the generation of speech. HMMs have lately been substituted by DNNs (Deep Neural Networks) (Shen 2017; Wang 2018), in which the training material is used to predict spectrograms which are then used to generate speech. In all these new approaches prosody has to be modeled separately and added at the beginning of the speech wave generation step. Prosody requires a new predictive model to be created again by means of a similar probabilistic approach. However, this is the most difficult component to recreate given the high level of variability present in natural language utterances. Google TTS system called TACOTRON 2 uses one such procedure but the conclusion of their approach is commented as follows: “We’d also like to develop techniques to select appropriate prosody or speaking style automatically from context, using, for example, the integration of natural language understanding with TTS.”

Alla sera
di Ugo Foscolo

Forse perchè della fatal quïete
Tu sei l'immago a me sì cara, vieni,
O Sera! E quando ti corteggian liete
Le nubi estive e i zeffiri sereni,

E quando dal nevoso aere inquiete
Tenebre, e lunghe, all'universo meni,
Sempre scendi invocata, e le secrete
Vie del mio cor soavemente tieni.

Vagar mi fai co' miei pensier su l'orme
Che vanno al nulla eterno; e intanto fugge
Questo reo tempo, e van con lui le torme

Delle cure, onde meco egli si strugge;
E mentre io guardo la tua pace, dorme
Quello spirto guerrier ch'entro mi rugge.

In morte del fratello Giovanni
di Ugo Foscolo

Un dì, s'io non andrò sempre fuggendo
Di gente in gente, mi vedrai seduto
Su la tua pietra, o fratel mio, gemendo
Il fior de' tuoi gentili anni caduto:

La madre or sol, suo dì tardo traendo,
Parla di me col tuo cenere muto:
Ma io deluse a voi le palme tendo;
E se da lunge i miei tetti saluto,

Sento gli avversi Numi, e le secrete
Cure che al viver tuo furon tempesta;
E prego anch'io nel tuo porto quiete:

Questo di tanta speme oggi mi resta!
Straniere genti, l'ossa mie rendete
Allora al petto della madre mesta.

Figure 2: Two poems by Ugo Foscolo annotated with spots difficult to read for TTS

In other words, deep neural networks are at pains predicting new unseen text and the accompanying speech waveforms without deep semantic knowledge. This is what SPARSAR system for poetry analysing and reading has been working on in the past twenty years or so (Delmonte 1981a; Delmonte 1981b; Delmonte et al. 1984). In order to show where deficiencies in current best TTS[1] are, we have chosen two Italian poems which have been read by demos on the web. Here above are the poems by Ugo Foscolo where we marked particularly difficult spots at the level both of word stress position and correct intonational contours. These poems have been chosen because they contain two specific phenomena: the first one are truncated words (fatal → fatale; guerrier → guerriero; corteggian → corteggiano; fratel → fratello; fior → fiore; sol → sola; viver → vivere; furon → furono). The other are enjambments – indicated with double arrows if within the same stanza, otherwise with a broken line when at stanza boundaries - requiring intonation to move from line-end to the following line and to the following stanza, when needed.

None of the systems we used has been able to pronounce truncated words correctly. This is reasonable seen that they don't belong to the current lexicon of nowadays Italian and that they are not using rule-based systems. However, also intonational contours have been badly mistaken: Nuance TTS has an internal rule to consider every line as end-stopped so that intonational contours always decrease at line-end. All enjambed lines are thus mistaken. The same happens with the other TTS "fromtexttospeech". The one by IBM manages to bridge the intonational continuation contour over more lines simply because it is based on the presence of punctuation marks. But expressivity is thus hampered.

[1] The public TTS demos we used are available at the following links: https://text-to-speech-demo.ng.bluemix.net/ - it produces an mp3 file which is always called TRANSCRIPT.mp3; https://www.nuance.com/omni-channel-customer-engagement/voice-and-ivr/text-to-speech.html - this has been perhaps derived from Italian TTS system Loquendo; http://www.fromtexttospeech.com/ - it produced an mp3 file which has a 8 number filename

2. *SPARSAR* – an Expressive TTS Reader

SPARSAR (Delmonte, 2015; Delmonte 2016) produces a deep analysis of each poem at different levels: it works at sentence level at first, then at verse level and finally at stanza level (see Figure 3 below). The structure of the system is organized as follows: the input text is processed at first at syntactic and semantic level and grammatical functions are evaluated. Then the poem is translated into a phonetic form preserving its visual structure and its subdivision into verses and stanzas. Phonetically translated words are associated to mean duration values taking into account position in the word and stress. At the end of the analysis of the poem, the system can measure the following parameters: mean verse length in terms of msec. and in number of feet. The latter is derived by a verse representation into metrical structure. Another important component of the analysis of rhythm is constituted by the algorithm that measures and evaluates rhyme schemes at stanza level and then the overall rhyming structure at poem level. In addition, the system has access to a restricted list of typical pragmatically marked phrases and expressions that are used to convey specific discourse function and speech acts and need specialized intonational contours.

We use the word "expressivity" in a specific general manner which includes sensible and sensitive reading that can only be achieved once a complete syntactic and semantic analysis has been provided to the TTS manager (Montaño, 2013; Saheer, 2013). Levels of intervention of syntactic-semantic and pragmatic knowledge include:

- syntactic heads which are quantified expressions
- syntactic heads which are preverbal SUBJects
- syntactic constituents that starts and ends an interrogative or an exclamative sentence
- distinguish realis from irrealis mood;
- distinguish deontic modality including imperative, hortative, optative, deliberative, jussive, precative, prohibitive, propositive, volitive, desiderative, imprecative, directive and necessitative etc.
- distinguish epistemic modality including assumptive, deductive, dubitative, alethic, inferential, speculative etc.
- any sentence or phrase which is recognized as a formulaic or frozen expression with specific pragmatic content
- subordinate clauses with inverted linear order; distinguishing causal from hypotheticals and purpose complex sentences
- distinguishing parentheticals from appositives and unrestricted relatives
- Discourse Structure to tell satellite and dependent clauses from main
- Discourse Structure to check for Discourse Moves - Up, Down and Parallel
- Discourse Relations to tell Foreground Relations from Backgrounds
- Topic structure to tell the introduction of a new Topic or simply a Change at relational level.

Current TTS are characterized by a total lack of expressivity. They only take into account information coming from punctuation and in some cases, from tagging. This hampers the possibility to capture the great majority of structures listed above. In particular, comma is a highly ambiguous punctuation mark with a whole set of different functions which are associated with specific intonational contours, and require semantic and discourse level knowledge to disentangle ambiguity. In general, question and exclamative marks are always used to modify the prosody of the previous word, which is clearly insufficient to reproduce such pragmatically marked utterances.

Few expressive speech synthesizers are tuned to specific domains and are unable to generalize. They usually convey specific emotional content linked to a list of phrases or short

utterances. In particular, Montaño et al. (2013) present an analysis of storytelling discourse modes and narrative situations, highlighting the great variability of speech modes characterized by changes in rhythm, pause lengths, variation of pitch and intensity and adding emotion to the voice in specific situations. Synthesis of specific expressive sounds is easy and can be extended easily to whole phrases by unit selection methods. This can be done by collecting separate databases for different emotions as suggested by Lakshmi et al. (2013).

2.1. The Module for Syntax and Semantics

The system uses a modified version of *VENSES*, a semantically oriented NLP pipeline (Delmonte et al. 2005). It is accompanied by a module that works at sentence level and produces a whole set of analysis both at quantitative, syntactic and semantic level. As regards syntax, the system makes available chunks and dependency structures. Then the system introduces semantics both in the version of a classifier and by isolating verbal complex in order to verify propositional properties, like presence of negation, to compute factuality from a crosscheck with modality, aspectuality – that is derived from the lexica – and tense. On the other hand, the classifier has two different tasks: separating concrete from abstract nouns, identifying highly ambiguous from singleton concepts (from number of possible meanings from WordNet and other similar repositories). Eventually, the system carries out a sentiment analysis of the poem, thus contributing a three-way classification: neutral, negative, positive that can be used as a powerful tool for prosodically related purposes.

Systems that can produce an appropriate semantic representation for a TTS are not many at an international level but they can be traced from the results of a Shared Task organized by members of SigSem and are listed here below in the corresponding webpage http://www.sigsem.org/w/index.php?title=STEP_2008_shared_task:_comparing_semantic_representations. State of the art semantic systems are based on different theories and representations, but the final aim of the workshop was reaching a consensus on what constituted a reasonably complete semantic representation. Semantics in our case not only refers to predicate-argument structure, negation scope, quantified structures, anaphora resolution and other similar items. It is referred essentially to a propositional level analysis, which is the basis for discourse structure and discourse semantics contained in discourse relations. It also paves the way for a deep sentiment or affective analysis of every utterance, which alone can take into account the various contributions that may come from syntactic structures like NPs and APs where affectively marked words may be contained. Their contribution needs to be computed in a strictly compositional manner with respect to the meaning associated to the main verb, where negation may be lexically expressed or simply lexically incorporated in the verb meaning itself.

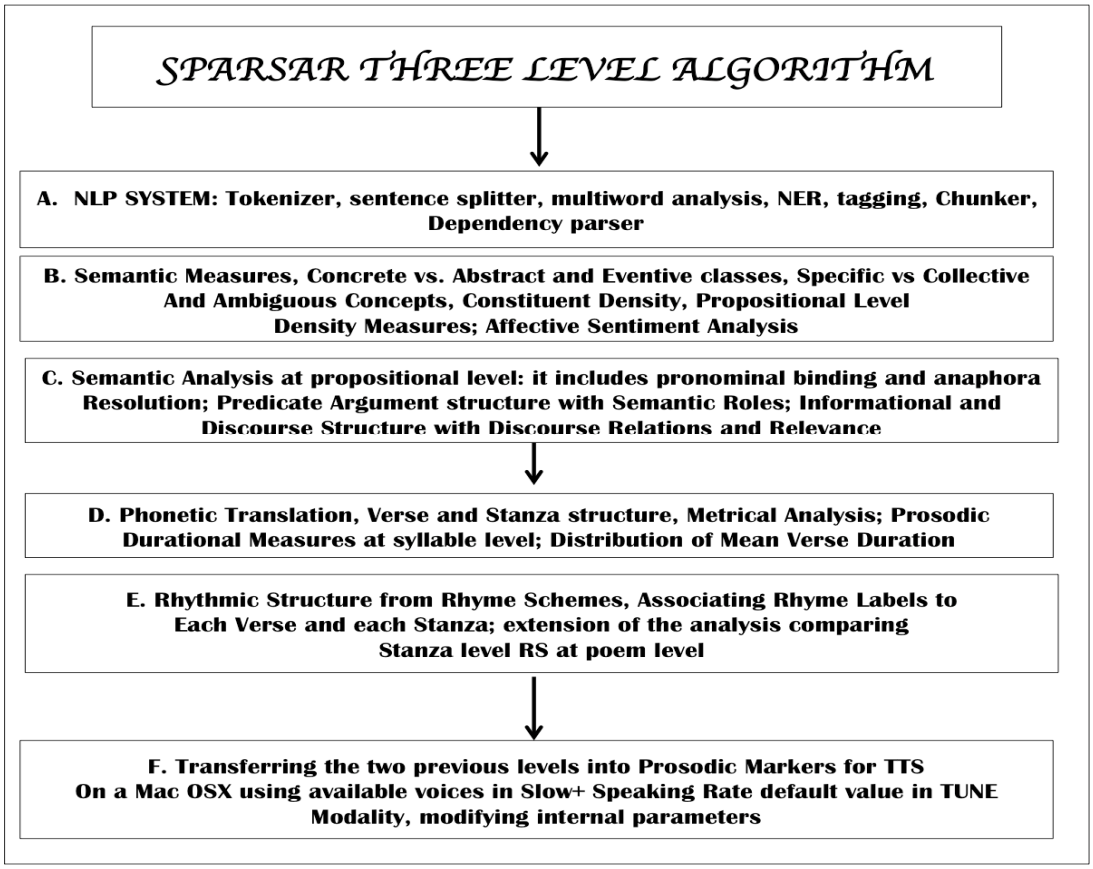


Figure 3: Architecture of *SPARSAR* with main pipeline organized into three levels

In Fig. 4 below we show the architecture of the deep system for semantic and pragmatic processing, in which phonetics, prosodics and NLP are deeply interwoven. The system does low level analyses before semantic modules are activated, that is tokenization, sentence splitting, multiword creation from a large lexical database. Then chunking and syntactic constituency parsing which is done using a rule-based recursive transition network: the parser works in a cascaded recursive way to include higher syntactic structures up to sentence and complex sentence level. These structures are then passed to the first semantic mapping algorithm that looks for subcategorization frames in the lexica made available for English, including VerbNet, FrameNet, WordNet and a proprietor lexicon of some 10K entries, with most frequent verbs, adjectives and nouns, containing also a detailed classification of all grammatical or function words. This mapping is done following LFG principles (Bresnan, 1982; Bresnan, 2001), where c-structure is mapped onto f-structure thus obeying uniqueness, completeness and coherence. The output of this mapping is a rich dependency structure, which contains information related also to implicit arguments, i.e. subjects of infinitivals, participials and gerundives. LFG representation also has a semantic role associated to each grammatical function, which is used to identify the syntactic head lemma uniquely in the sentence. Finally, it takes care of long distance dependencies for relative and interrogative clauses. When fully coherent and complete predicate argument structures have been built, pronominal binding and anaphora resolution algorithms are fired. Coreferential processed are activated at the semantic level: they include a centering algorithm for topic instantiation and memorization that we do using a three-place

stack containing a Main Topic, a Secondary Topic and a Potential Topic. In order to become a Main Topic, a Potential Topic must be reiterated. Discourse Level computation is done at propositional level by building a vector of features associated to the main verb of each clause. They include information about tense, aspect, negation, adverbial modifiers, modality. These features are then filtered through a set of rules which have the task to classify a proposition as either objective/subjective, factual/nonfactual, foreground/background. In addition, every lexical predicate is evaluated with respect to a class of discourse relations. Eventually, discourse structure is built, according to criteria of clause dependency where a clause can be classified either as coordinate or subordinate. We have a set of four different moves to associate to each clause: root, down, level, up.

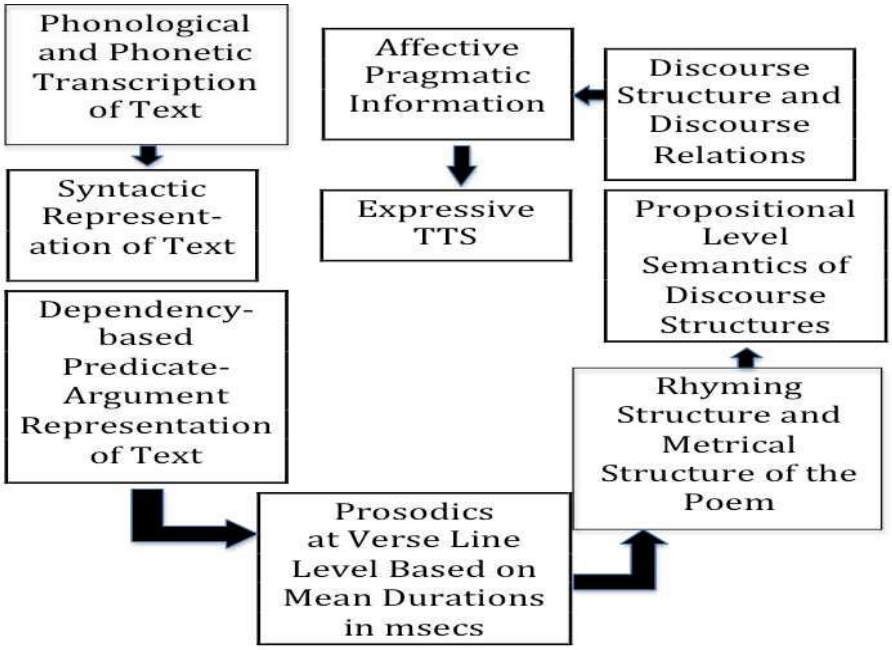


Figure 4: Interconnections between Syntactic-Semantic representations and Phonetic-Prosodic Rules

2.2. The Modules for Phonetics and Prosody

The second module is a rule-based system that converts graphemes of each poem into phonetic characters, it divides words into stressed/unstressed syllables and computes rhyming schemes at line and stanza level. To this end it uses grapheme to phoneme translations made available by different sources, amounting to some 500K entries, and include CMU dictionary[2], MRC Psycholinguistic Database[3], Celex Database (Baayen et al. 1995), plus a proprietor database made of some 20,000 entries. Out of vocabulary words are computed by means of a prosodic parser implemented in a previous project (Bacalu & Delmonte, 1999a, b) containing a

[2] It is available online at http://www.speech.cs.cmu.edu/cgi-bin/cmudict/.
[3] Previously, data for POS were merged in from a different dictionary (MRC Psycholinguistic Database, http://lcb.unc.edu/software/multimrc/multimrc.zip, which uses British English pronunciation)

big pronunciation dictionary which covers 170,000 entries approximately. Besides the need to cover the majority of grapheme to phoneme conversions by the use of appropriate dictionaries, remaining problems to be solved are related to ambiguous homographs like “import” (verb) and “import” (noun) and are treated on the basis of their lexical category derived from previous tagging. Eventually there is always a certain number of Out Of Vocabulary Words (OOVW). The simplest case is constituted by differences in spelling determined by British vs. American pronunciation. This is taken care of by a dictionary of graphemic correspondences. However, whenever the word is not found the system proceeds by morphological decomposition, splitting at first the word from its prefix and if that still does not work, its derivational suffix. As a last resource, an orthographically based version of the same dictionary is used to try and match the longest possible string in coincidence with current OOVW. Then the remaining portion of word is dealt with by guessing its morphological nature, and if that fails a grapheme-to-phoneme parser is used. Here below are some of the OOVWs that have been reconstructed by means of the recovery strategy explained above: we indicated each example by showing the input word rejected by the dictionary lookup, then the word found by subtraction and the final output obtained by recomposition:

% wayfarer → [wayfare-[w_ey1f_eh1_r_r]],
% gangrened → [gangrene-[g_ae1_nr_ah0_n_d]],
% krog → [krog-g_r_aa1_g]
% copperplate → [copper-k_aa1_p_er_p_l_ey1_t],
% splendor → [splendour-[s_p_l_eh1_n_d_er]],
% filmy → [film-f_ih1_l_miy],
% seraphic → seraphine--> [s_e_r_a_ph_iy1_k]
% unstarred → [starred-[ah_n_s_t_aa1_r_d]]

Other words we had to reconstruct are: shrive, slipstream, fossicking, unplotted, corpuscle, thither, wraiths, etc. In some cases, the problem that made the system fail was the presence of a syllable which was not available in our database of syllable durations, *VESD* (Bacalu & Delmonte 1999a; Bacalu & Delmonte, 1999b). This problem has been coped with by manually inserting the missing syllable and by computing its duration from the component phonemes, or from the closest similar syllable available in the database. We only had to add 12 new syllables for a set of approximately 1000 poems that the system computed.

The system has no limitation on type of poetic and rhetoric devices, however it is dependent on language: Italian line verse requires a certain number of beats and metric accents which are different from the ones contained in an English iambic pentameter. Rules implemented can demote or promote word-stress on a certain syllable depending on selected language, line-level syllable length and contextual information. This includes knowledge about a word being part of a dependency structure either as dependent or as head. A peculiar feature of the system is the use of prosodic measures of syllable durations in msec: in this way, a theoretic prosodic measure for each line and stanza can be produced, using mean durational values associated to stressed/ unstressed syllables. This index is called “prosodic-phonetic density index”, because it contains count of phones plus count of theoretic durations: the index is intended to characterize the real speakable and audible consistency of each line of the poem. A statistic is issued at different levels to evaluate distributional properties in terms of standard deviations, skewness and kurtosis. The final output of the system is a parameterized version of the poem which is then read aloud by the speech synthesis system: parameters are generated taking into account all previous analysis including sentiment or affective analysis and discourse structure, with the aim to produce an expressive reading (but see below for an example).

As R. Tsur (2012) comments in his introduction to his book, iambic pentameter has to be treated as an abstract pattern and no strict boundary can be established. The majority of famous English poets of the past, while using iambic pentameter have introduced violations, which in some cases – as for Milton's Paradise Lost – constitute the majority of verse patterns. Instead, the prosodic nature of the English language needs to be addressed, at first. English is a stress-timed language as opposed to Spanish or Italian which are syllable-timed languages. As a consequence, what really matters in the evaluation of iambic pentameters is the existence of a certain number of beats – 5 in normal cases, but also 4 in deviant ones. Unstressed syllables can number higher, as for instance in the case of exceptional feminine rhyme or double rhyme, which consists of a foot made of a stressed and an unstressed syllable (very common in Italian), ending the line - this is also used by Greene et al. (2010) to loosen the strict iambic model. These variations are made to derive from elementary two-syllable feet, the iamb, the trochee, the spondee, the pyrrhic. According to the author, these variations are not casual, they are all motivated by the higher syntactic-semantic structure of the phrase. So there can be variations as long as they are constrained by a meaningful phrase structure.

In our system, in order to allow for variations in the metrical structure of any line, we operate on the basis of syntactic dependency and have a stress demotion rule to decide whether to demote stress on the basis of contextual information. The rule states that word stress can be demoted in dependents in adjacency with their head, in case they are monosyllabic words. In addition, we also have a promotion rule that promotes function words which require word stress. This applies typically to ambiguously tagged words, like "there", which can be used as expletive pronoun in preverbal position, and be unstressed; but it can also be used as locative adverb, in that case in postverbal position, and be stressed. For all these ambiguous cases, but also for homographs not homophones, tagging and syntactic information is paramount.

Our rule system tries to avoid stress clashes and prohibits sequences of three stressed/three unstressed syllables, unless the line syntactic-semantic structure allows it to be interpreted otherwise. Generally speaking, prepositions and auxiliary verbs may be promoted; articles and pronouns never. An important feature of English vs. Italian is length of words in terms of syllables. As may be easily gathered, English words have a high percentage of one-syllable words when compared to Italian which on the contrary has a high percentage of 3/4-syllable words.

2.3. Computing Metrical Structure and Rhyming Scheme

Any poem can be characterized by its rhythm which is also revealing of the poet's peculiar style. In turn, the poem's rhythm is based mainly on two elements: meter, that is distribution of stressed and unstressed syllables in the verse, presence of rhyming and other poetic devices like alliteration, assonance, consonance, enjambments, etc. which contribute to poetic form at stanza level. This level is combined then with syntax and semantics to produce the adequate breath-groups and consequent subdivision: these will usually coincide with line stop words, but they may continue to the following line by means of enjambments.

What is paramount in our description of rhythm, is the use of the acoustic parameter of duration. The use of acoustic duration allows our system to produce a model of a poetry reader that we implement by speech synthesis. The use of objective prosodic rhythmic and stylistic features, allows us to compare similar poems of the same poet and of different poets both prosodically and metrically. To this aim we assume that syllable acoustic identity changes as a function of three parameters: internal structure in terms of onset and rhyme which is

characterized by number of consonants, consonant clusters, vowel or diphthong; position in the word, whether beginning, end or middle; primary stress, secondary stress or unstressed.

As discussed above - see Figure 1, the analysis starts by translating every poem into its phonetic form. After processing the whole poem on a line by line basis and having produced all phonemic transcription, the system looks for poetic devices. Here assonances, consonances, alliterations and rhymes are analysed and then evaluated. Here metrical structure is computed, that is the alternation of beats: this is done by considering all function or grammatical words which are monosyllabic as unstressed. In particular, "0" is associated to all unstressed syllables, and a value of "1" to all stressed syllables, thus including both primary and secondary stressed syllables. Syllable building is a discovery process starting from longest possible phone sequences to shortest one. This is done heuristically trying to match pseudo syllables with the syllable list. Matching may fail and will then result in a new syllable which has not been previously met. The assumption is that any syllable inventory will be deficient, and will never be sufficient to cover the whole spectrum of syllables available in the English language. For this reason, a certain number of phonological rules has been introduced in order to account for any new syllable that may appear. To produce the prosodic model, mean durational values are considered. Whenever possible, also positional and stress values are selected. Also syntactic information is taken advantage of, which computed separately to highlight chunks' heads as produced by bottomup parser. In that case, stressed syllables take maximum duration values. Dependent words on the contrary are "demoted" and take minimum duration values.

Durations are then collected at stanza level and a statistic is produced. Metrical structure is used to evaluate its distribution in the poem by means of statistical measures. As a final consideration, we discovered that even in the same poem it is not always possible to find that all lines have identical number of syllables, identical number of metrical feet and identical metrical verse structure. If we consider the sequence "01" as representing the typical iambic foot, and the iambic pentameter as the typical verse metre of English poetry, there is no poem strictly respecting it in our analyses. On the contrary we found trochees, "10", dactyls, "100", anapaests, "001"and spondees, "11". At the end of the computation, the system is used to measure two important indices: "mean verse length" and "mean verse length in no. of feet" that is mean metrical structure.

Additional measures that we are now able to produce are related to rhyming devices. Since we consider important taking into account structural internal rhyming schemes and their persistence in the poem the algorithm makes available additional data derived from two additional components: word repetition and rhyme repetition at stanza level. Sometimes also "refrain" may apply, that is the repetition of an entire line of verse. Rhyming schemes together with metrical length, are the strongest parameters to consider when assessing similarity between two poems.

Eventually the internal structure of metrical devices used by the poet can be reconstructed: in some cases, also stanza repetition at poem level may apply. We then use this information as a multiplier. The final score is tripled in case of structural persistence of more than one rhyming scheme; it is doubled for one repeated rhyme scheme. With no rhyming scheme there will be no increase in the linear count of poetic and rhyming devices. To create the rhyming scheme couples of rhyming lines are searched by trying a match recursively of each final phonetic word with the following ones, starting from the closest to the one that is further apart. Each time both rhyming words and their distance are registered. In the following pass, the actual final line numbers are reconstructed and then an indexed list of couples, Line Number-Rhyming Line for all the lines is produced, including stanza boundaries. Eventually, alphabetic labels to each rhyming verse starting from A to Z. A simple alphabetic incremental mechanism updates the

rhyme label. This may go beyond the limits of the alphabet itself and in that case, double letters are used.

2.4. Transforming Poetic and Semantic Data into Parameters for Speech Synthesis

Here below the sequence of rules where linguistic, prosodic and poetic triggers are called by parameters' creation modules divided up into four separate prosodically defined levels. As can be easily gathered, the rules require linguistic knowledge at all levels of analysis, which has been produced by previous steps of analysis. Prosodic Markers Induction works at different levels:

Level 1. PAUSE INSERTION

- ✓ a word is a syntactic head (either at constituency or dependency level)
- ✓ a word is a quantifier, a quantified adverbial, or it marks the beginning of a quantified expression
- ✓ a word is a discourse marker and defines the beginning of a subordinate clause
- ✓ a word is a SUBJect head

Level 2. RHYTHMIC CONTROL

- ✓ the title
- ✓ first and last line of the poem
- ✓ a word marks the end of a line and is (not) followed by punctuation
- ✓ a word is the first word of a line and coincides with a new stanza, and is preceded by punctuation
- ✓ word stress demotion for words dependent on a following head

Level 3. INTONATIONAL CONTROL

- ✓ a word is the first/last word of an exclamative sentence
- ✓ a word is the first/last word of an interrogative sentence and is (not) the question constituent
- ✓ a line is part of a sentence which is a frozen or formulaic expression with specific pragmatic content and is exceptionally phonetically encoded
- ✓ a line is part of a sentence that introduces a New Topic, a Change, a Foreground Relevance content as computed by the Semantics in Discourse Relations
- ✓ a line is part of a sentence and is dependent in Discourse Structure and its Move is Down or Same Level

Level 4. PHONETIC SEGMENTAL CONTROL

- ✓ a word is one of a list of phonetically spelled out words which are wrongly computed by the TTS and need direct input
- ✓ an expression or utterance is a frozen or formulaic expression and requires specialized intonational and phonetic direct input

We will apply these rules to the poem *Edge* by Sylvia Plath, written one week before her death by suicide. Here below the text of the poem with my Italian translation. The peculiarity of this poem is shown by the way in which stanzas have been created: the poem is organized into three macro stanzas with the first two made up by four micro couplets and the last one made up by two micro couplets. In addition, every second line in the first eight couplets is strongly enjambed with the first line of the following couplet. This requires a very careful control of both intonation and rhythm which is not achieved by commercial TTS. The only TTS that manages to read the poem decently is the one by IBM which however overlooks completely the subdivision into couplets and into macro stanzas.

Edge

By Sylvia Plath 1963

The woman is perfected.
Her dead

Body wears the smile of accomplishment,
The illusion of a Greek necessity

Flows in the scrolls of her toga,
Her bare

Feet seem to be saying:
We have come so far, it is over.

Each dead child coiled, a white serpent,
One at each little

Pitcher of milk, now empty.
She has folded

Them back into her body as petals
Of a rose close when the garden

Stiffens and odors bleed
From the sweet, deep throats of the night flower.

The moon has nothing to be sad about,
Staring from her hood of bone.

She is used to this sort of thing.
Her blacks crackle and drag.

Limite/Bordo/Orlo/Margine

La donna ora è perfetta.
Il suo corpo

morto indossa il sorriso della compiutezza,
l'illusione di una greca necessità

sgorga nelle pieghe della sua toga,
i suoi nudi

piedi sembrano dire:
siamo arrivati fin qui, è la fine.

Ogni bimbo morto acciambellato, serpente bianco,
ciascuno a una piccola

brocca di latte, ora vuota.
Lei li ha riavvolti

di nuovo nel suo corpo come i petali
di una rosa si chiudono quando il giardino

s'intorbidisce e i profumi sanguinano
dalle dolci, profonde gole del fiore notturno.

La luna non ha nulla di cui esser triste,
osserva fisso dal suo cappuccio d'osso.

E' assuefatta a questo tipo di cose.
Il suo nero sipario striscia e scricchiola.

Figure 5: Sylvia Plath's Edge and its Italian translation

Themes developed in the macro stanzas are related but different: the first theme is the woman, while the second one are dead children – begotten by the woman - which are then transformed by strongly symbolic metaphors and synaesthesia in flowers and gardens. The two final couplets recall the woman in the beginning of the poem again by means of strongly symbolic metaphors, this time suggested by the analogy woman and moon.

3. Computing Rhyme Violations in Shakespeare's Sonnets

In this final section a recent version of *SPARSAR* will be presented which is dedicated to Elizabethan English poetry. I will discuss problems one has to cope with in dealing with Early Modern English (hence EME) texts, which are written in a peculiar format, the one of Elizabethan and in particular, Shakespearean sonnets. In addition, I will explain what consequences ensue from the presence of diachronic variants on the recital by TTS. Shakespeare wrote the sonnets before his death in 1616 in a period in which Early Modern English was still in use but starting to undergo substantial changes. We don't know precisely how words were pronounced but we know for sure what an Elizabethan sonnet should sound like due to rhyming conventions which were very strict at the time, poetry being mainly rehearsed and only occasionally available on printed paper. In this way, specific words may be deemed to have a different sound from the current one. Since a lexicon of a language should encompass the basic sound of a word, together with its grammatical, syntactic and semantic properties, we will be concerned with the way in which a system for poetry recital through TTS - including *SPARSAR*, should account for these variations. I will be using previous TTS systems available on the web, including this time Google's SPEAK in which input could be organized using SSML (Speech Synthesis MarkUp Language[4]).

Shakespeare wrote 154 sonnets, with a total of 18,283 tokens and 3085 types - vocabulary richness of 16.87% being in line with the best poets. These numbers to certify that words were chosen very carefully to suit the variety of topics narrated in the sonnets, but also to abide the constraints imposed by rhyming structure. Also number of Hapax and Rare Words (indicating the union of Hapax, Dis and TrisLegomena) corresponds to average values for other poets, respectively to 56%, the first type and 79% the second one. It is important to show quantitative data of word usage in poetry, because the choice of any word in a sonnet has to be made with great care, not only for syntactic and semantic reasons, but also because sound properties of words play an enormous role in the economy of poetic devices. Elizabethan and Shakespearean sonnets are organised on the basis of iambic pentameter and have alternate rhymes distributed into three quatrains and a final couplet which is in perfect rhyme. This applies to all sonnets, excluding no. 99 - which is made up of 15 lines organised in chained rhyme -, and no. 126 which only has 12 lines. All lines are pentameter feet sometimes with feminine ending, i.e. with one additional syllable (11 not 10), again excluding no. 145 which is made of lines with tetrameter feet. In many cases, however, respecting iambic pentameter is guaranteed by the presence of a great number of contractions, and the opposite prohibition of elision of end of word unstressed syllable.

3.1. Coping with Contractions and Violations

Contractions are present in great number in the sonnets. Computing them requires reconstructing their original complete corresponding word form in order to be able to match it to the lexicon or simply derive the lemma through morphological processing. This is essentially due to the fact that they are not predictable and must be analysed individually. Each type of contraction has a different manner to reconstruct the basic wordform. In order to understand and reconstruct it correctly, each contraction must go through a recovery procedure of the lemma.

[4] A complete manual is avaible on the web at this link: https://www.w3.org/TR/speech-synthesis11/.

We have found 821 contractions in the collection, where 255 are cases of genitive ‘s, and 167 are cases of past tense/participle ‘d. The remaining cases are organised as follows:

> SUFFIXES attached at word end, for example [‘s, ‘d, ‘n, ‘st, ‘t, (putt’st)]
> PREFIXES elided at word beginning, for example [‘fore, ‘gainst, ‘tis, ‘twixt, ‘greeing]
> INFIXES made by consonant elision inside the word [o’er, ne’er, bett’ring, whate’er, sland’ring, whoe’er, o’ercharg’d, ‘rous]

Finally, as an example of a stanza where elisions are not allowed, consider the following one taken from sonnet no.46. In particular, consider the two verbs “impanelled” and “determined” at the end of line, which contribute to the overall iambic meter with two feet each. Using *SPARSAR* to analyse the sonnets has allowed us to evaluate all poetic devices including rhyme schemes. In this way, we discovered more than 100 violations to the typical sonnet rhyme scheme. This theme has been discussed and reported in papers and also on a website - http://originalpronunciation.com/ - by linguist David Crystal.

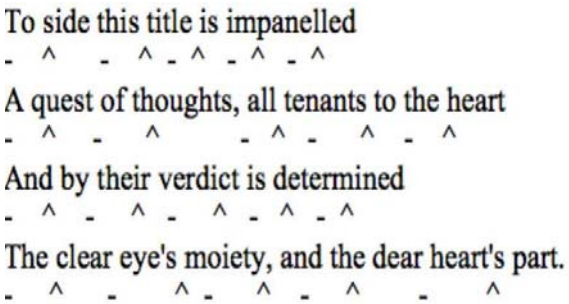


Figure 6: Annotated stanza with iambic pentameters from sonnet 46

His position is however very disputable: even though he is convinced that the original pronunciation may be induced by rhyme schemes, he then describes the phenomenon claiming it is merely phonological. To this aim, he organized the violations into separate phonological classes differentiating them by vowel-related transformations and consonant-related ones, adding those phenomena that merely displace word stress from one syllable to another. In presence of alternate pronunciation for the same word, he claims these are exceptions, giving examples of similar cases in today’s pronunciation for words that do not change meaning but only sounds (“again”, “says”, “often”, “schedule”, etc.).

From a careful study of the first 100 words in Crystal’s list – which includes also words coming from the plays -, one can easily understand that the phenomenon is not describable simply using phonological rules: the same vowel in stressed position is assigned to an indefinite number of transformation disregarding its context. No reference is made to neighbouring syllables nor to stress patterns. On the contrary, we believe variants to be lexically determined. To support this approach, we referred ourselves to the book published in 1889 by the famous historical phonologist Wilhelm Vietor, by the title “A Shakespeare Phonology”. To represent the phenomenon, we will also be using a phonological representation but then we will report words that have undergone transformation in the system and how this has been realized.

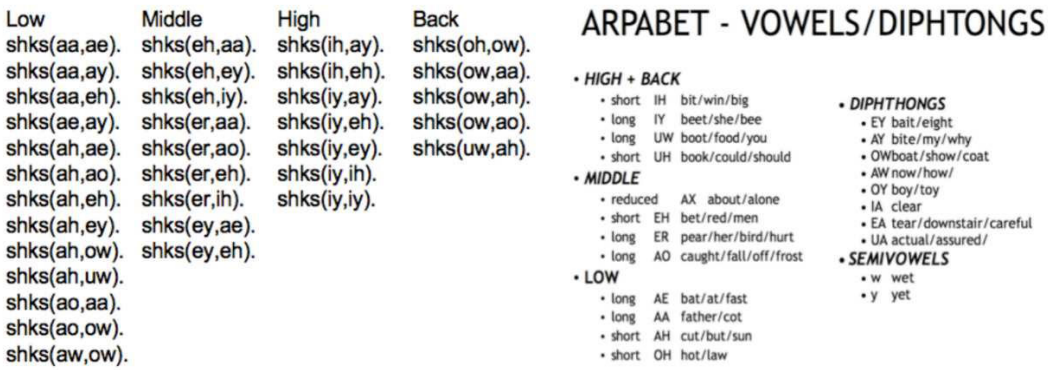


Figure 7: Transformations applied to vowels subdivided by phonological classes and Arpabet phonetic alphabet

Here above a table of all vowels and their transformations organized, as suggested by D. Crystal, according to their phonological class, using Arpabet as phonetic alphabet. Vowels and their transformations are represented in a Prolog compliant notation shks(Vov1,Vow2), where Vow1 is the input current vowel of a given word in contemporary phonetic dictionaries, and Vow2 is the old corresponding sound. As can be easily noticed, variants are applied often to the same vowel, Vow1, thus clearly showing the impossibility to apply rules of any type. Words involved in the transformation are listed below in the excerpt taken from the internal lexicon made available to *SPARSAR*. As can be easily noticed, variants are related to both stress position and to consonant sounds.

Lexicon 1.
shks(despised,d_ih2_s_p_ay1_s_t,ay1,ay1)
shks(dignity,d_ih2_g_n_ah_t_iy1,iy1,ay1).
shks(gravity,g_r_ae2_v_ah_t_iy1,iy1,ay1).
shks(history,hh_ih2_s_t_er_iy1,iy1,ay1).
shks(injuries,ih2_n_jh_er_iy1_z,iy1,iy1).
shks(jealousy,jh_eh2_l_ah_s_iy1,iy1,ay1).
shks(jollity,jh_aa2_l_t_iy1,iy1,ay1).
shks(majesty,m_ae2_jh_ah_s_t_iy1,iy1,ay1).
shks(memory,m_eh2_m_er_iy1,iy1,ay1).
shks(nothing,n_ah1_t_ih_ng,ah1,ow1).

The entries of this portion of the lexicon reported under Lexicon 1, are organized into four slots: first slot is the word as it appears in the text; slot 2, contains the phonetic translation available in CMU dictionary which however has been modified to sound Elizabethan in more than one place in some cases. Not all entries are doubly modified, however. The first entry, DESPISED, only reports the change in the final devoiced syllable which has been changed into S_T from Z_D. DIGNITY on the contrary, has both stress position and final syllable vowel sound modified into DIGNI'TAY, etc. In the second list reported below under Lexicon 2, we see on the contrary words which have a double non conservative variant and can thus be pronounced in two manners according to rhyming constraints – see below.

Lexicon 2.
shks(moan,m_ow1_n,ow1,aa1).
shks(moan,m_ow1_n,ow1,ao1).

```
shks(gone,g_ao1_n,ao1,aa1).
shks(gone,g_ao1_n,ao1,ow1).
shks(gone,g_ao1_n,ao1,ao1).
```

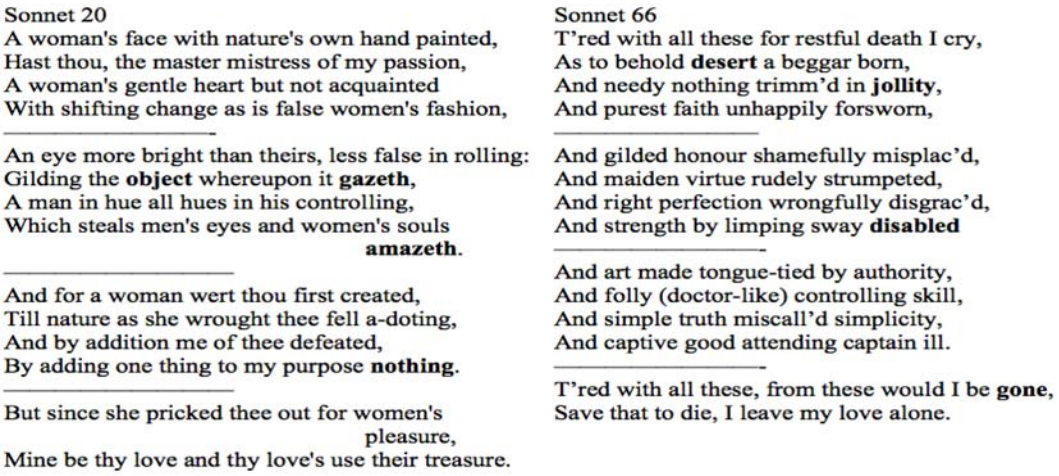
It is now clear that variants need to interact with information coming from the rhyming algorithm that alone can judge whether the given word, usually at line end - but the word can also be elsewhere, has to undergo the transformation or not. The lexicon in our case has not been built manually but automatically, by taking into account all rhyming violations and transcribing the pair of word at line end on a file. The algorithm searches couple of words in alternate lines inside the same stanza and whenever the rhyme is not respected it writes the pair in output. Take for instance the pair LOVE/PROVE, in that order in alternate lines within the same stanza: in this case it is the first word that has to be pronounced like the second. The order is decided by the lexicon: LOVE is included in the lexicon with the rule for its transformation, PROVE is not. In some other cases it is the second word that is modified by the first one, as in CRY/JOLLITY, again the criterion for one vs the other choice is determined by the lexicon.

So basically, a new rhyming algorithm was created which is triggered by the choice to analyse Elizabethan sonnets. The algorithm modifies the phonetic form of a given word in case it is recognized as one of the words belonging to the dictionary, and at the same time if it is violating rhyming rules. The algorithm is organized as follows:

- Browse the list of end-of-line words
    - Current word (CW) IS contained in the Lexicon of Phonological Variants (LPV)?
        - Check the rhyming pair by searching downward within the same stanza the alternate rhyming line
            - If the word has the same stressed vowel predicted in the rule associated to the previous word
            - Iterate if needed
            - MODIFY the current word (CW)
        - Else SKIP
    - Current word (CW) is NOT contained in LPV?
        - Check the rhyming pair searching upward within the same stanza the alternate rhyming line
            - If the word is contained in LPV
            - Check to verify if the stressed vowel predicted in the rule is the same of the stressed vowel associated to CW
            - Iterate if needed
            - MODIFY the second rhyming word
        - Else SKIP

3.2. Making a TTS pronounce the correct word sound

Once the text has been correctly transformed into the phonetic representation adequate to reproduce the expected rhyming scheme of the sonnet, the reading process has to be likewise instructed. Using a commercially available TTS on the Mac allows the reader to take advantage of its ability which however in this case would be of no use. Input to the TTS should be the text transformed into phonetic forms derived from the conversion algorithm.



Sonnet 20
A woman's face with nature's own hand painted,
Hast thou, the master mistress of my passion,
A woman's gentle heart but not acquainted
With shifting change as is false women's fashion,
An eye more bright than theirs, less false in rolling:
Gilding the **object** whereupon it **gazeth**,
A man in hue all hues in his controlling,
Which steals men's eyes and women's souls **amazeth**.
And for a woman wert thou first created,
Till nature as she wrought thee fell a-doting,
And by addition me of thee defeated,
By adding one thing to my purpose **nothing**.
But since she pricked thee out for women's pleasure,
Mine be thy love and thy love's use their treasure.

Sonnet 66
T'red with all these for restful death I cry,
As to behold **desert** a beggar born,
And needy nothing trimm'd in **jollity**,
And purest faith unhappily forsworn,
And gilded honour shamefully misplac'd,
And maiden virtue rudely strumpeted,
And right perfection wrongfully disgrac'd,
And strength by limping sway **disabled**
And art made tongue-tied by authority,
And folly (doctor-like) controlling skill,
And simple truth miscall'd simplicity,
And captive good attending captain ill.
T'red with all these, from these would I be **gone**,
Save that to die, I leave my love alone.

Figure 8: Sonnet 20 and 66 showing in bold difficult to read words for a commercial TTS

I marked with italics and bold those words that the TTS pronounced wrongly and that required intervention from SPARSAR; and I marked in bold those words that were included in the specialised dictionary for EME pronunciation and that were required to rhyme with a previous or following end-of-line word. In addition, notice the presence of typical EME words ending with TH (**gazeth**, **amazeth**) or simply T (**wert**) for third person present tense agreement, none of which can be correctly pronounced by the TTS – neither the Mac, nor the Google one, nor even Nuance's. Then consider the two words **OBJECT** and **DESERT**: the first one had word stress positioned by the TTS on the second syllable whereas the current use as a noun required stress on first syllable. The second word was pronounced with its current meaning and stress on first syllable: on the contrary, its usage in EME was quite different and derived from the verb DESERVE. The pronunciation is then similar to "dessert", with stress on second syllable. Finally, the word DISABLED which has to rhyme with STRUMPETED: both words should be treated as trisyllabic, but in order to get that pronunciation "disabled" has to be modified. This word is contained in a line which requires an additional syllable to be added, so that the correct pronunciation required by the metrical iambic pentameter has to become "**disabel-e-d**". This second transformation is induced by metrical count which is operated independently of rhyming concerns, and will switch the rule for syllable insertion.

Here below the transformation of sonnet 66 into the parameterized form to be used as input to Macintosh speech synthesis with the voice of Alex. It can be copy pasted into a txt file produced by TextEdit where the function Speak is made available and get it read by the computer. Instructions on how to use prosodic parameters under OS X is made available by Apple in the Speech Synthesis Programming Guide issued in 2006 which however is now superseded by the new voices produced by Nuance that don't allow a complete interface with those parameters. In particular, intonational variations are no longer allowed. It can only be used with the old voices, including Alex. As can be easily seen, main parameters available are these:

- pauses indicated by parameter SLNC (dubbing silence);
- volume indicated by parameter VLM;
- speaking rate or speed by parameter RATE;
- speech pitch or intonation by parameter PBAS (dubbing baseline pitch);
- a speech command to express emphatic emotion EMPH;

- to input phonetic version of a word or phrase [[inpt PHON]];
- to end phonetic input [[inpt TEXT]];

Apart from Volume, all other parameters are expressed as numerical values. In particular Pitch can use the range of real values from 1000 to 127.000, where 60.000 represents middle value on a conventional piano and it corresponds to 261.625 Hz (ibid.: 16); Rate has an average typical conversational speed is 180 and it can be pushed up to 500;

> [[pbas 38.000; rate 160; volm +1.5]] ***sonnet 66 by William_Shakespeare*** . [[slnc 400]],[[rset 0]] [[pbas 44.000; rate 145; volm +1.5]] ***tired with*** [[rate 120; volm +1.5]] ***all these*** [[pbas 40.000; rate 120; volm +1.5]][[slnc 300]] ***for restful death I*** [[pbas 38.000; rate 130; volm +1.5]] ***cry*** [[slnc 200]] ***, as to behold*** [[inpt PHON]] ***dAXzEHrt*** [[inpt TEXT]] ***a beggar*** [[pbas 38.000; rate 130; volm +1.5]] ***born*** [[slnc 200]] , [[slnc 100]] ***and needy nothing trimmed in*** [[pbas 38.000; rate 130; volm +1.5]][[inpt PHON]] ***JOHllIHt1AY*** [[inpt TEXT]][[slnc 200]] , [[slnc 100]] ***and purest faith unhappily*** [[pbas 38.000; rate 130; volm +1.5]] ***forsworn*** [[slnc 200]] ***,*** [[slnc 100]] ***and gilded honour shamefully*** [[pbas 38.000; rate 130; volm +1.5]] ***misplaced*** [[slnc 200]] ***,*** [[slnc 100]] ***and maiden virtue rudely*** [[pbas 38.000; rate 130; volm +1.5]] ***strumpeted*** [[slnc 200]] , [[slnc 100]] ***and right perfection wrongfully*** [[pbas 38.000; rate 130; volm +1.5]] ***disgraced*** [[slnc 200]] , [[slnc 100]] ***and strength by limping sway*** [[pbas 36.000; rate 120; volm -0.2]] [[inpt PHON]] ***dIHs1EYbAXl-lEHd*** [[inpt TEXT]] [[slnc 200]] ***and art made tongue tied by*** [[pbas 38.000; rate 130; volm +1.5]] ***authority*** [[slnc 200]] ***,*** [[slnc 100]] ***and folly (doctor like) controlling*** [[pbas 38.000; rate 130; volm +1.5]] ***skill*** [[slnc 200]] ***,*** [[slnc 100]] ***and simple truth miscalled*** [[pbas 38.000; rate 130; volm +1.5]] ***simplicity*** [[slnc 200]] ***,*** [[slnc 100]] ***and captive good attending captain*** [[pbas 38.000; rate 130; volm +1.5]] ***ill*** [[slnc 200]] .
> [[pbas 54.000; rate 155; volm +1.5]] ***tired with*** [[rate 120; volm +1.5]] ***all these*** [[pbas 40.000; rate 120; volm +1.5]][[slnc 300]] ***, from these*** [[rate 120; volm +1.5]] ***would*** [[slnc 100]] ***I be*** [[pbas 38.000; rate 130; volm +1.5]] ***gone*** [[slnc 200]] ***, save that to die , I leave my love*** [[pbas 38.000; rate 130; volm +1.5]] ***alone*** [[slnc 200]]

4. Conclusion

In this article we presented work carried out to allow a computer read aloud English and Italian Poetry and recently added modules that deal specifically with Elizabethan poetry and Early Middle English pronunciation. We have started by focussing on TTS, i.e. the TextToSpeech module which is at the heart of any speech synthesizer and is responsible for naturalness and expressivity. Eventually this can only be achieved by introducing NLU capabilities in the TTS, i.e. Natural Language Understanding ability to allow the system to instruct the Prosodic Manager if any producing appropriate parameters. Prosody is responsible for correct placement and length of pauses, adequate speaking rate and consequent ability to modulate it when needed (speaking slower or faster), adequate intensity of voice volume to reproduce emotions, and what's more important intonational contours. Modelling prosody automatically by means of statistical approaches is not achievable due to the high level of variability of speaking modes by different speakers and diversity introduced by different genres of text being read. Thus *SPARSAR* has been created to fill the gap and does it correctly. A number of poems have been presented and commented and eventually a fully parameterized sonnet by Shakespeare is presented so that it can be used to check the ability of the system. Related demo mp3 files are available at dedicated website sparsar.wordpress.com.

REFERENCES

Alan W. Black, et al., (2011), *New Parameterization for Emotional Speech Synthesis, Final Report for NPESS team, CSLP Johns Hopkins Summer Workshop 2011*, http://www.clsp.jhu.edu/workshops/ws11/groups/npess.

Alan W. Black, et al., (2012), *Articulatory Features For Expressive Speech Synthesis*, in *Proceeding of ICASSP 2012*, IEEE, 4005-4008.

Archana Balyan, S. S. Agrawal, Amita Dev, (2013), *Speech Synthesis: A Review*, *International Journal of Engineering Research & Technology (IJERT)*, Vol. 2 Issue 6, pp. 57-75.

Baayen R. H., R. Piepenbrock, and L. Gulikers, (1995), *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium.

Bacalu C., Delmonte R., (1999a), *Prosodic Modeling for Syllable Structures from the VESD - Venice English Syllable Database*, in *Atti 9° Convegno GFS-AIA*, Venezia.

Bacalu C., Delmonte R., (1999b), *Prosodic Modeling for Speech Recognition*, in *Atti del Workshop AI*IA - "Elaborazione del Linguaggio e Riconoscimento del Parlato"*, IRST Trento, pp. 45-55.

Bresnan J. (ed.), (1982), *The Mental Representation of Grammatical Relations*, The MIT Press, Cambridge MA.

Delmonte R. (1981a), *An Automatic Unrestricted Tex-to-Speech Prosodic Translator*, in *Atti del Convegno Annuale A.I.C.A.*, Pavia, pp. 1075-83.

Delmonte R. (1981b), *Automatic Word-Stress Patterns Assignment by Rules: a Computer Program for Standard Italian*, in *Proc. IV F.A.S.E. Symposium*, 1, ESA, Roma, pp. 153-156.

Delmonte R., (2008), *Speech Synthesis for Language Tutoring Systems*, in V. Melissa Holland & F. Pete Fisher (eds.), (2008), *The Path of Speech Technologies in Computer Assisted Language Learning*, Routledge - Taylor and Francis Group, New York, pp. 123-150.

Delmonte R., (2015), *Visualizing Poetry with SPARSAR - Poetic Maps from Poetic Content*, in *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, Denver, Colorado, Association for Computational Linguistics, pp. 68–78.

Delmonte R., (2016), *Expressivity in TTS from Semantics and Pragmatics*, in Vayra, M., Avesani, C. & Tamburini F. (eds.) (2016), *Il farsi e disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio/Language acquisition and language loss. Acquisition, change and disorders of the language sound structure*, Milano, AISV. pp. 407-427.

Delmonte R., et al., (2005), *VENSES – a Linguistically-Based System for Semantic Evaluation*, in J. Quiñonero-Candela et al.(eds.), *Machine Learning Challenges*. LNCS, Springer, Berlin, pp. 344-371.

Delmonte R., G.A. Mian, G. Tisato, (1984), *A Text-to-Speech System for the Synthesis of Italian*, in *Proceedings of ICASSP'84*, San Diego(Cal), pp. 291-294.

Delmonte, R., Francesco Stiffoni, (2011), *Using Speech Synthesis to Simulate an Interlanguage and Learn the Italian Lexicon, SLATE_2011*, in *Proceedings Workshop "Speech and Language Technology in Education"*, Venezia, ISCA Archives, pp. 25-28, downloadable at https://www.isca-students.org/archive/slate_2011/papers/sl11_025.pdf.

Greene E., T. Bodrumlu, K. Knight, (2010), *Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation*, in *Proceedings of the 2010 Conference on EMNLP*, pp. 524–533.

Jonathan Shen and Ruoming Pang, (2017), *Tacotron 2: Generating Human-like Speech from Text*, *Google AI Blog*, https://ai.googleblog.com/2017/12/tacotron-2-generating-human-like-speech.html.

Montaño, Raúl, Francesc Alías, Josep Ferrer, (2013), *Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis*, in *8th ISCA Speech Synthesis Workshop*, pp. 171-176.

Rajeswari K. C., Uma Maheswari P., (2012), *Prosody Modeling Techniques for Text-to-Speech Synthesis Systems - A Survey*, in *International Journal of Computer Applications* (0975 – 8887) Vol. 39, No.16, pp. 8-11.

Saheer, Lakshmi, Blaise Potard – IDIAP, Martigny, (2013), *Understanding Factors in Emotion Perception*, at *8° Speech Synthesis Workshop*, pp. 59-64.

Sproat, R. (ed.), (1997), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Dordrecht, Kluwer Academic.

T. Delić, S. Suzić, M. Sečujski, D. Pekar, (2017), *Multi-style Statistical Parametric TTS*, in *Proceedings Digital Speech and image processing* (DOGS 2017), pp. 5–8.

Tsur Reuven, (2012), *Poetic Rhythm: Structure and Performance: An Empirical Study in Cognitive Poetics*, Sussex Academic Press, pp. 472.

Yuxuan Wang, RJ Skerry-Ryan, (2018), *Expressive Speech Synthesis with Tacotron, Google AI Blog*, https://ai.googleblog.com/2018/03/expressive-speech-synthesis-with.html.

RODOLFO DELMONTE • has been associate professor of Computational Linguistics at Ca Foscari University in Venice since 1987 and is now retired. He has over 200 publications in international journals and conference proceedings, including 8 books. He has been member of scientific committees and chair of international conferences, invited speaker in Europe, Australia and in USA. Expert for national research councils like SSHR in Canada, ANR and AERES in France, Scientific Research Fund (FWO) in Belgium, and the European Commission in Brussel. Teaching courses and seminars in summers schools in Romania and Bulgaria, in Paris, Besançon, San Sebastian, and for a longer period invited professor at the University UTD in Dallas. He has organized 27 prestigious workshops and conferences in Venice. He and his team participated in semantically related international Challenges organized by NIST and ACL in USA, as well as by EVALITA in Italy with his symbolic linguistically based system called Getaruns. Semantics and pragmatics has always been his focus of interest, including research themes like irony, sarcasm in political commentaries, literary narratives and poetry lately including Shakespeare's Sonnets. In the last five years he has decided to dedicate himself to poetry and TTS which were his first research interests and has produced SPARSAR a system that reads English poetry preserving meaning and contributing emotions.

E-MAIL • delmont@unive.it