MuLAS: A Framework For Automatically Building Multi-Tier Corpora

Sérgio Paulo, Luís C. Oliveira

L^{2}F - Spoken Language Systems Lab
INESC-ID/IST
{spaulo,lco}@l2f.inesc-id.pt

Abstract

The **Mu**lti-**L**evel **A**lignment **S**ystem (MuLAS) is the L2F tool for building multi-tier speech corpora with reduced or no human intervention at all. MuLAS automatically combines information coming from external speech annotations, human or machine-generated, with the text-based utterance descriptions that it creates, in order to build more reliable and complete descriptions of the spoken utterances.

This paper presents our methods for multi-tier annotation synchronization, which lie behind the MuLAS operation. Such methods have allowed us to expand the building of multi-tier corpora to new languages without spending too much effort. MuLAS has been successfully applied to the building of multi-tier corpora for speech synthesis in American and British English, European Portuguese and German. Natural prosody generation has benefited from MuLAS, too, since prosodic models can be derived from corpora built by MuLAS.

Index Terms: Multi-tier corpora, speech annotation, multi-level alignment, temporal overlap.

1. Introduction

The dream of using any human being's voice in a text-to-speech system, even when the voice owner is not alive anymore, became true in the last fifteen years with the unit-selection synthesizers. However, speech production is a complex cognitive task involving a hierarchical structure of execution that extends from the production of syntactically and semantically organized sentences down to the production of the phone sequence. The multi-layer nature of speech raises the need for annotations at multiple levels of granularity. Setting explicit relations among multiple annotation levels is not a subject of minor relevance, since it accounts for much of the co-occurring and inter-related phenomena lying behind the speech production process. The availability of multi-tier corpora arises as a requisite for an accurate modeling of that process.

Even the phonetic segmentation, which is more concerned with the detection and location of phonetic segments [2, 4, 7], can benefit from this sort of corpora, as higher level contexts can be used to set more accurate predictions about the most likely word pronunciations.

Since manual annotation of speech is time-consuming, automatic methods have been successfully applied to perform this task at some specific levels[1] of the utterance's linguistic representation. Even though sometimes automatic annotations are not reliable enough to be used without manual verification, the availability of such annotations speeds up the overall process to a large extent [8].

[1]Mostly for the phonetic level, although some methods have been used for prosodic annotation, too.

When it comes to build multi-tier corpora, the most frequently used approach has consisted in creating graphical tools [1, 3, 6], designed in order to minimize the time needed for building such corpora by human labelers. Building these corpora is still highly dependent on human effort, which sets an economic barrier that prevents many (minor) languages from being deeply studied [5].

We propose a method to build multi-tier corpora with minimal human intervention. While the set of tiers addressed in this work is primarily tailored to speech synthesis and natural prosody generation, our algorithm can be modified to deal with different tiers, so that other issues can be addressed, too. With this algorithm, multi-tier corpora can be built by human labelers and automatic tools, simultaneously. Moreover, the time-consuming human setting of inter-tier connections through graphical tools, [1, 3, 6], can be avoided, since the algorithm is able to do it automatically.

MuLAS enables a user to derive further annotations from those provided to the system. Therefore, the user can decide not to provide annotations at some linguistic levels, and accept the estimates set by the system based on the whole utterance description, created in the meantime. For example, the annotation of the disjuncture level between consecutive words (word break indexes) is not commonly available. Thus, in such a case, MuLAS predicts fairly accurate estimates of word break indexes by relating the punctuation marks with the phonetic information inherited from the respective word phonemes. The utterances are subsequently re-phrased according to that word-level feature estimates.

This paper is organized as follows: section 2, provides a general overview of the MuLAS operation; section 3, presents the method that we propose for setting relations between the predicted and observed phone sequences; section 4, presents our approach for synchronizing the utterances' higher level descriptions, and section 5 is reserved to the conclusions.

2. System Overview

Fig. 1 is a schematic representation of MuLAS operation, which builds on three major modules:

- Text Analysis (TexA)
- Two-Stage Predicted/Observed Phone Sequence Alignment (ToSPOPA)
- Higher-level Synchronization (HLS)

The *TexA* module is responsible for the creation of a *standard* representation of spoken utterances based on their orthographic transcriptions, only. Such a representation comprises a set of predictions spanning from the phrase and intonation levels down to the fundamental frequency level. In order to make it easier to expand the use of MuLAS to new languages, we

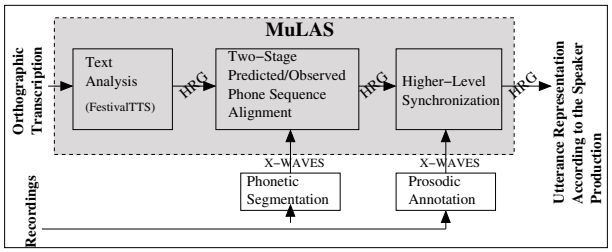

Figure 1: *Schematic representation of the MuLAS operation.*

decided that the *TexA* module should be based on the Festival Speech Synthesis System [9], which is a multi-lingual system with a growing number of supported languages. Therefore, the Heterogeneous Relation Graphs (HRG) [10] formalism, used in the Festival system, was also adopted as our system's logical data structure.

The *ToSPOPA* module sets a relationship between the utterance's phone-level representations predicted from the text and observed from the phonetic segmentation of the speech signal. This module establishes how predicted phones (speaker-independent) are mapped onto the phones produced by the speaker.

The *HLS* module performs local modifications on the higher level utterance descriptions predicted by *TexA* in order to remove any mismatch between such predictions and the available speech signal annotations.[2]

2.1. Predicted Tiers

The *TexA* module creates exactly the same utterance representation as *Festival* does while synthesizing arbitrary text, except the waveform generation related tiers and features. As the predicted descriptions get closer to the speech signal, they become more detailed, thus more sensitive to speaker variations. For instance, it is more likely that a speech synthesizer can accurately predict the prosodic phrasing than the fundamental frequency values of the recorded speech signals. This way, some predicted descriptions (tiers) will be kept almost untouched, while others will be completely re-loaded.

A corpus built by MuLAS relies on four major utterance descriptions: phonetic/phonological, prosodic phrasing, intonation and grammatical tagging. The grammatical tagging is the tier that lies farthest from the speech signal, since it is completely defined by the written text. Thus, it does not depend upon the speech produced by the speaker at all. Moreover, automatic grammatical tagging has been reported to be more than ninety-percent accurate in English, thus, it is likely that few mistakes are made at this level.

Both the phonetic/phonological and intonation level descriptions are associated with the temporal information of the recorded speech signals, thus, are more sensitive to speaker-specific reading strategies. In order to avoid overwriting those utterance description levels, which would be a hard hit on the automatization of the corpus building procedure, the adaptation of the predicted representations must be carried out by means of local modifications of those representations. MuLAS followed the latter approach for automatically creating realistic utterance representations.

[2]word breaks, intonation events, F0 values, etc.

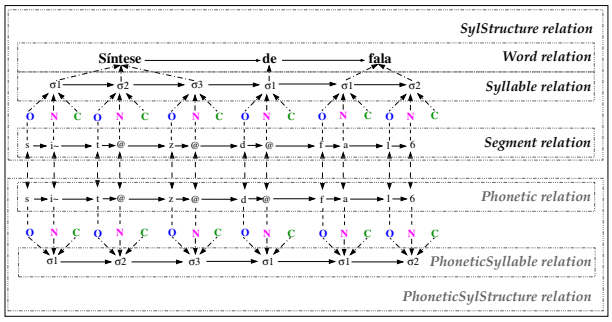

Figure 2: *Schematic representation of the Festival relations used for representing the phonetic-phonological data.*

3. Phonetic/Phonological Tiers

The phonetic/phonological tiers build on a set of *Festival* utterance relations[3] along with some others specially created for this purpose, as depicted in Fig. 2. The black- and gray-colored relations of that figure are the default relations of the *Festival* utterance representation and the newly created relations, respectively.

Table 1: *Festival's newly created relations.*

Relation	Definition
PhoneticSyllable	A simple list of items representing the phonetic syllables arisen from the phonetic sequence produced by the speaker.
Phonetic	A simple list of items accounting for the phones uttered by the speaker."
PhoneticSylStructure	A list of tree structures over the items in the Phonetic and PhoneticSyllable relation items.

In MuLAS framework, *Syllable* and *Segment* relations were moved slightly away from the phonetic production of the speaker. That is, *Syllable* and *Segment* relations represent streams of phonological syllables[4] and their constitutive phonemes, respectively. Therefore, such relations account for the early linguistic analysis of the text prompts, keeping themselves from addressing the post-lexical phonological processes, which are more prone to the speaker variability. Let this relation set be the *phonologically-motivated* utterance representation. The speaker-dependent phonetic issues - like the phonetic sequence that the speaker actually produces - build on the results of the phonetic segmentation and are addressed by the newly created *Phonetic*, *PhoneticSyllable* and *PhoneticSylStructure* relations, whose definitions are shown in Table 1. From now on, these three relations are referred as the *phonetically-motivated* utterance representation.

Thus, we have both the phonologically- and phonetically-motivated utterance representations in hands, and must create a relation among them in order to bridge the gap between the predicted representations and the recorded utterance. Such a relation consists in a *multi-linear* structure, since it must allow for an item at the *Segment* relation to be associated to one or more items at the *Phonetic* relation and *vice-versa*.

[3]Word, Syllable, Segment, SylStructure, IntEvent, Intonation, Target and Phrase relations. Further information about this issue is available at http://festvox.org

[4]Arisen from the isolated pronunciations of *Word* relation items

In order to address that problem, a novel two-stage algorithm was implemented in MuLAS. Firstly, a set of one-to-one relationships is established between the predicted and observed phones, based on an alignment between the two phone sequences, let it be called *one-to-one phone mapping procedure.* That alignment is comprised of phone substitutions, insertions and deletions.[5]

In order to obtain a bi-directional mapping between the phone sequences, *inserted* and *deleted* phones must not keep floating. That is, they must be associated with, at least, one phone of the other phonetic sequence. Such associations are set in the second stage of this algorithm, which can be called *one-to-many phone mapping.*

3.1. One-to-one Phone Mapping Procedure

The minimal set of operations needed to transform a predicted phone sequence into the sequence of observed phones is found by computing the alignment between both sequences. A prior restriction must be made in order to compute meaningful alignments: every phone substitution that lacks a minimal theoretical support is forbidden. Table 2 shows the set of allowed phone substitutions of our algorithm.

Table 2: *Allowed phone substitutions.*

Phone belonging to:	Can be substituted by:
Vowels	a vowel or a glide
Glides	a vowel or a glide
Consonants	a consonant
Silences	a silence segment

After setting such constraints, insertion, deletion and substitution costs are calculated according to the procedure described in [11]. The phone sequence alignment is also described in that paper.

With a phone sequence alignment in hands, association links between the *Segment* and *Phonetic* relation items can be finally set. By the previously described reasons, such associations should be made by means of a *multi-linear structure*, thus the alignment computed just before is not enough to establish a complete relationship between Phonetic and Segment relation items. Moreover, due to the inability of *Festival* to represent such data structures, two hierarchic relations are defined in order to account for phone associations in a single direction each: *PhonemeToPhone* and *PhoneToPhoneme*. The former relation aims at setting association links for *inserted* phones, while the latter performs the same task for the *deleted*.

Both the *PhonemeToPhone* and *PhoneToPhoneme* relations consist in tree lists. The roots and leaves of the *PhonemeToPhone* relation trees are in the *Segment* and *Phonetic* relations, respectively. Although the *PhoneToPhoneme* relation builds on the same two relations, they play opposite roles in this case. Only those phones that were unmodified or substituted can be roots of trees in the respective relations. Therefore, each one of them is set a connection to the other relation's phone that it was aligned with: *one-to-one phone mapping.* Inserted and deleted phone connections are not set at this stage.

[5]When the phone sequence alignment comprises recorded phones that are not aligned with any of predicted phones, we say that an insertion took place. When it happens the other way round, a deletion occurred.

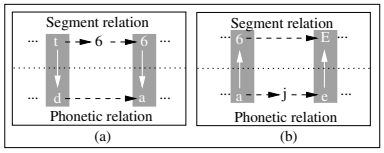

Figure 3: *Anchor points for phone deletions (a) and insertions (b).*

3.2. One-to-many Phone Mapping Procedure

Inserted phones belong to the *Phonetic* relation. In order to be properly accessed from the higher levels of the utterance representation, they must be associated to the predicted phonetic segments - lying at the *Segment* relation - that gave rise to them. Such associations are set in the scope of the *PhonemeToPhone* relation, so that several observed phones can be associated with any single predicted phone. When it comes to deal with phone deletions, everything processes the other way round. That is, one has to associate the deleted phone (lying at the *Segment* relation) to the observed phone (at the *Phonetic relation*) that resulted from a possible fusion with any of its neighbors. Those associations occur within the *PhoneToPhoneme* relation. At this stage, associations are rule-based, as described later on in this paper.

3.2.1. Anchor Points

Anchor points are those phones, either at the *Segment* or *Phonetic* relations, whose association links were set at the *one-to-one phone mapping stage*. The elements inside the gray-colored boxes of Figure 3 are the anchor points resulting from the *one-to-one phone mapping* procedure.

Except when an insertion or a deletion occur at the beginning or the end of the sequence, in which a single association possibility exists, one has always two possible *anchor points* for an inserted and deleted phone association.

Thus, an inserted or deleted phone associates either to the left or to the right anchor. That ambiguity is solved by means of a set of rules, accounting for the phonological processes that can have caused such insertions or deletions to come up.

3.2.2. Setting association links for deleted and inserted phones

Insertions and deletions can occur by two distinct reasons: post-lexical phonological processes or distinct word pronunciations. This part is language-dependent, and rules can be used to decide whether the *floating* phones shall be associated to the left or to the right anchor. In absence of any matching rule for a phone context, an inserted phone is associated to the left anchor, while a deleted phone is associated to the right anchor.

Now, how are such contexts set? The contexts used to address phone insertions and deletions are set according to expressions (1) and (2), respectively,

$$Pn(l)\_On(l) + On(ph) + On(r)\_Pn(r) \quad (1)$$

$$Pn(l)\_On(l) + Pn(ph) + On(r)\_Pn(r) \quad (2)$$

where $Pn(l)$, $On(l)$, $On(ph)$, $Pn(ph)$, $On(r)$, $Pn(r)$ stand for the left anchor's predicted phone name, left anchor's observed phone name, inserted phone name, deleted phone name, right anchor's observed phone name and right anchor's predicted phone name, respectively. Symbols "_" and "+" of those expressions represent an association of phones in different relations (levels) and a phone boundary, respectively.

Whenever the default association is not meaningful, rules can be used to modify such associations. Rules are provided

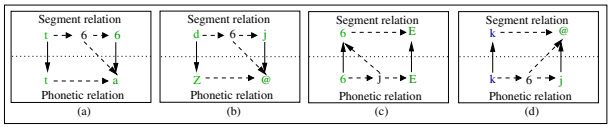

Figure 4: *(a) Deletion raised by a phonological process; (b) Deletion raised by a mismatching word pronunciation; (c) Insertion arisen raised by a phonological process; (d) Insertion raised by a mismatching word pronunciation.*

as shown in Table 3, in which the ".*" symbol stands for any string, as a rule is selected by means of a regular expression matching between its context and the predicted (or observed) phone context.[6] Although the set of rules shown here account for the insertion or deletion of a single phone, there is no restriction on this issue, and further rules can be provided to the system in order to account for multiple consecutive deletions and insertions.

Table 3: *Association rules for phone insertions and deletions.*

Insertion context	Anchor	Rule Number
"6"_"6"+"j"+"E"_"E"	left	1
.*_.*+"6"+"j"_"@"	right	2
Deletion context	**Anchor**	**Rule Number**
"t"_"t"+"6"+"@"_"j"	right	3
.*_.*+"6"+"a"_"6"	left	4

Fig. 4 shows four distinct contexts that MuLAS can be asked to address while building a corpus. The cases shown in Fig. 4 a, b, c and d, triggered the rules number 4, 3, 1 and 2, respectively.

After all these operations (performed within the ToSPOPA module), the phonetic-level descriptions of the speaker's productions is connected to the remaining utterance structure. Moreover, such connections enable the utterance structure to have access to the time information of the speech signal.

4. Higher-Level Synchronization

The temporal information can now flow from the *Phonetic* to *Segment* relation through the *PhoneToPhoneme* and *PhonemeToPhone* relations. After setting the time boundaries of the *Segment* relation items, the *Syllable* and *Word* time boundaries are automatically known, since they are inherited from the *Segment* relation items. Thus, the HLS module can start its operation.

If any prosodic annotation of the speech signal is available, it can be used to refresh the utterance descriptions at that level. By using the temporal inclusion as a criterion intonation events can be assigned to the respective syllables. Word break indexes are also updated at this stage, either by using the respective annotations, when available, or by deriving word breaks from the co-occurrence of phonetic phenomena - like duration stretching, silences - and punctuation marks at the word level. In order to achieve more realistic descriptions, the utterances are rephrased according to the newly created word break indexes.

The resulting utterance structure is a much more reliable description of the recorded utterance. Thus, MuLAS could build that description without too much human intervention.

5. Conclusions

This paper presented a new approach for building multi-tier speech corpora by automatic means. Moreover, it presented a method that allows for the creation of an explicit relation between a *standard* realization of the text prompts and the speaker-specific production. With such a relation, one can derive models to account for the post-lexical phonological phenomena. Besides, we can create inventories of context-dependent word pronunciations, which can play an important role in both speech synthesis and recognition systems. Given the automatic nature of the approach proposed here, we believe that MuLAS can be a powerful tool for studying many less-resourced languages.

[6]The phone instances shown in this paper belong to the SAMPA phone set for European Portuguese, described at https://www.l2f.inesc-id.pt/ lco/ptsam/ptsam.pdf

6. Acknowledgements

This work was partially funded by PRIME National Project TECNOVOZ number 03/165 and by the FCT project POSC/PLP/58697/2004. It was also partially supported by European Community (EC) in the scope of the eCIRCUS project IST-4-027656-STP.

7. References

[1] Barras, C., Geoffrois, E., Wu, Z., Liberman, M., "Transcriber: Development and use of a tool for assisting speech corpora production.", Speech Communication, 33, 5–22, 2001.

[2] Carvalho, P., Trancoso, I., Oliveira, L. C., "Automatic segment alignment for concatenative speech synthesis in portuguese." Proc. of the 10th Portuguese Conference on Pattern Recognition, RECPAD'98, 221–226, 1998.

[3] Cassidy, S., Harrington, J., "Multi-level annotation in the EMU speech database management system." Speech Communication, 33, 61–77, 2001.

[4] Cox, S., Brady, R., Jackson, P., "Techniques for accurate automatic annotation of speech waveforms." Proc. of the International Conference on Spoken Language (ICSLP '98), 1947–1950, 1998.

[5] Greenberg, S., "Strategies for automatic multi-tier annotation of spoken language corpora." Proc. of the 8th European Conference on Speech Communication and Technology (Interspeech 2003), 45–48, 2003.

[6] McKelvie, D., Isard, A., Mengel, A., Moller, M. B., Grosse, M., Klein, M., "The MATE Workbench - an annotation tool for xml coded speech corpora." Speech Communication 33, 97–112, 2001.

[7] Paulo, S., Oliveira, L. C., "DTW-based phonetic alignment using multiple acoustic features." Proc. of the 8th European Conference on Speech Communication and Technology (Interspeech 2003), 309–312, 2003.

[8] Syrdal, A. K., Hirschberg, J., McGory, J., Beckman, M., "Automatic tobi prediction and alignment to speed manual labeling of prosody." Speech Communication, 33, 135–151, 2001.

[9] Black, A. W., Taylor, P., and Caley, R., "The Festival Speech Synthesis System documentation". 2002

[10] Taylor, P., Black, A. W., and Caley, R., "Heterogeneous relation graphs as a formalism for representing linguistic information." Speech Communication, 33, PAGES ,2001.

[11] Paulo, S., and Oliveira, L. C., "Multilevel Annotation of Speech Signals Using Weighted Finite State Transducers." Proc. of IEEE 2002 Workshop on Speech Synthesis, Santa Monica, California, 2002.