Multi-domain text classification for unit selection Text-to-Speech synthesis

Francesc Alías, **Ignasi Iriondo** and **Pere Barnola**

Enginyeria i Arquitectura La Salle. Universitat Ramon Lull
Pg. Bonanova 8, 08022 - Barcelona, Spain
{falias,iriondo,tm05122}@salleURL.edu

Abstract

This paper presents a new approach for designing a concatenative text-to-speech (TTS) system based on multi-domain unit selection. The method achieves good synthetic quality with reasonable computational cost for a general-purpose TTS system. The architecture of the multi-domain database and the text classification algorithm for domain assignment are the basis of the method. The performance of the adjusted text classification algorithm for the multi-domain TTS aim is analyzed in several encouraging experiments.

1 Introduction

Recently, concatenative Text-to-Speech (TTS) synthesis systems have move from diphone based approaches, with only one instance per unit, to *unit selection* based methods [1, 2, 3]. These methods are based on a large speech database allowing multiple instances per unit.

These corpus based methods have shown that natural-sounding synthetic speech can be achieved by minimizing the number of concatenation points and reducing the amount of prosodic modification, overcoming the drawbacks of the diphone based techniques. The unit selection module searches the database to obtain the set of units that best matches the target specifications (according to a cost function [1]) by means of dynamic programming (for instance, the Viterbi algorithm).

In fact, it is important to note that the database has to be designed to cover as much linguistic variability as possible for a particular language or domain [4]. The characterization of the database is still an on-going research issue [5]. Nevertheless, it is clear that the computational cost at synthesis time grows exponentially with the size of the database. Furthermore, in order to enhance the synthetic quality, the unit selection systems have been applied to restricted domains giving high quality in synthetic speech within domain [6].

To take advantage of the speech quality obtained with the limited-domain approach without discarding a general purpose system, we present a new TTS system based on a multi-domain structured database (see section 2). In this framework, a procedure for domain classification becomes indispensable, therefore, we have designed an automatic algorithm based on natural language processing (NLP) for text classification (see section 3). Hence, the sentences that compose the input text are assigned to the appropriate domain or set of domains. The accuracy of the algorithm is presented in different experiments that evaluate its performance as a text classifier. Moreover, after tuning the text classification algorithm for the TTS purpose, we present the improvements obtained by the multi-domain paradigm (see section 4). Finally, the conclusions of this work are discussed (see section 5).

2 Multi-domain unit selection TTS

The main application for the unit selection speech synthesis is a general-purpose TTS system (GP-TTS), which can produce *any* desired utterance from an input text [1, 2, 3]. Although the synthetic speech quality is usually very high, there are still bad synthesis examples in the unit selection GP-TTS conversion [6]. Therefore, in order to improve this issue, the unit selection process has been applied to limited domains (LD-TTS), achieving very high quality within those domains (see [5, 7] for a review).

Furthermore, it has been suggested that the synthetic quality of the TTS conversion heavily reflects the style and coverage of the recorded database [4, 8]. Thus, the synthetic quality of a GP-TTS system decreases when the target domain of the input text mismatches the coverage of the GP speech database [6, 9]. Chu et alt. [9] also give good reasons for improving a GP-TTS system. The work takes into account some domain adaptation to achieve natural speech.

Allowing for these ideas, we present a multi-domain TTS (MD-TTS) system (see Figure 1), in order to obtain high synthetic quality (like the LD-TTS approach) in a GP-TTS. The system, by means of text classification, assigns the input sentences to a domain or set of

ISBN 1-876346-48-5

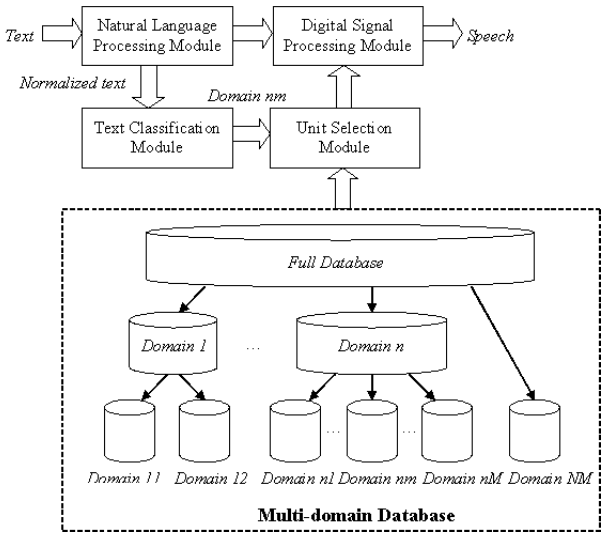

Figure 1: Block diagram of the proposed multi- domain unit selection TTS system.

domains ($D_{nm} \supseteq D_{NM}$) from the database (see section 3). It is important to note that while the vocabulary of every domain database will certainly be specialized, it has to be designed with the same phonetic and prosodic coverage of a GP-TTS, minimizing the OOV effect of word concatenation LD-TTS systems [7].

3 Text classification algorithm

Text classification (TC) can be modelled as a supervised learning task with the main goal of classifying a document into a set of predefined categories. TC methods are developed by the machine learning community within the artificial intelligence paradigm. There are many methods for TC, some of them are based on decision trees, neural nets (NN), Bayesian approaches, or Support-Vector Machines [10] among others.

These methods are often based on the *bag-of-words* approach, representing texts by means of a vector space model [11]. As a result, each document is represented using a vector of weights related to the occurrence of terms within the text, ignoring their relationships. Before the final term representation, some pre-processing strategies are applied for classification looking for relevant linguistic components (stop listing and stemming) and significant semantic features (dimensionality reduction)[10]. Neither strategies are applied in our TC algorithm, since our main goal is to classify the *full* sentence in the appropriate domain, paying more attention to the style of the text. As depicted in the Figure 1, the only applied pre-processing step is the typical NLP normalization module in the TTS conversion process (expanding numbers, dates, etc.).

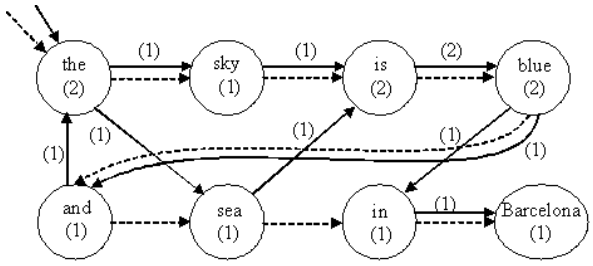

Figure 2: Weighted ARN obtained from the text *"The sky is blue and the sea is blue in Barcelona"*.

3.1 Associative Relation Networks

In the speech generation paradigm, the continuity and the style of speech are two key features to obtain a good synthetic speech quality. In this context, it is important to find an alternate representation for the typical vector-space model, taking into account word relations. This approach seems to be unfeasible due to the parameterization of the input text as a weighted collection of independent words.

In a first approximation, the proposed TC system represents the information of each domain by means of an Associative Relation Network (ARN) [12] defined as a weighted graph of connected nodes. In the ARN approach, patterns are made up of the terms which compose the input text, but also by their relationships to each other. The nodes of the graph represent words and their connections describe the co-occurrences with other words in the same text, including its direction (see Figure 2 for an example). As a result, this kind of network encodes the coarticulation effect at word boundaries and information about the structure of the text (*style*), which is an essential element in the multi-domain approach to unit selection TTS synthesis.

3.2 Modeling the input text

There are several proposed text parameterization schemes ranging from the simple boolean estimate to more sophisticated approaches. However, most of them ignore the order of occurrence of terms and their organization [10]. In order to take into account the lexical structure and the style of the text, within a real-time TC purpose, the considered parameters (*par*) are three:

- **Term frequency** (tf): the number of times that the word appears in the text [11], which is included in every node of the ARN (Figure 2).
- **Co-occurrence frequency** (cof): the number of times that two words take place together in the text, thus, the relations between words are considered [12]. In our approach, this parameter accounts for the connections of the text and it is represented in the ARN as the weight between two related nodes in the graph (see Figure 2).

ISBN 1-876346-48-5

- **Pattern length (pl)**: the number of consecutive words of the input text that appear successively in the ARN of the model, normalized by the total number of input words, thus, $[0 \leq pl \leq 1]$. It evaluates the membership of the input text into the model of the analyzed domain.

The standard *idf* parameter, i.e. the inverse document frequency [11], is not included in this first approach since it belongs to the semantic viewpoint for TC.

3.3 Similarity measures

Similarity measures are used to quantify the resemblance between the input text (t_k) and the domain model (D_{nm}). We have defined several distances for text classification in the MD-TTS context:

$$\vec{v} = \big(tf_1, cof_{11}, .., tf_i, cof_{i1}, .., cof_{ij}, .., tf_I, cof_{IJ}\big) \quad (1)$$

$$S_1\big(t_k, D_{nm}\big) = \frac{\vec{v}_{t_k} \cdot \vec{v}_{D_{nm}}}{\|\vec{v}_{t_k}\| \cdot \|\vec{v}_{D_{nm}}\|} \quad (2)$$

$$S_2\big(t_k, D_{nm}\big) = p\,l \cdot S_1\big(t_k, D_{nm}\big) \quad (3)$$

$$sigmoid(x) = \frac{1}{1 + e^{-\left(\frac{x-\mu}{\sigma}\right)}} \quad (4)$$

$$S_3\big(t_k, D_{nm}\big) = \left(p\,l \cdot sigmoid\Big(\sum_i tf_i^{t_k} - tf_i^{D_{nm}}\Big) \cdot sigmoid\Big(\sum_{i,j} cof_{ij}^{t_k} - cof_{ij}^{D_{nm}}\Big)\right)^{1/3} \quad (5)$$

First, an adaptation of the popular cosine similarity [11] is presented in Equation (2). This measure determines the angle between two vectors of parameters (tf_i and cof_{ij}, in Equation (1)) as a comparison distance. Secondly, the cosine is weighted by the $p\,l$ to emphasize the linguistic structure of text (see Equation (3)). Finally, we present a measure derived from the *sigmoid* function of the NN pattern (Equation (4)). The sigmoid is tuned by its activation threshold μ and its shape σ (a larger value making the curve flatter). This similarity is obtained as the geometric average of sigmoid differences in tf and cof, weighted by $p\,l$.

3.4 Multi-domain model generation

Every model of a domain D_{nm} is an ARN that has been implemented as a classic decision tree with a superimposed indexing array (dashed line in Figure 2). The generation process can be supervised, starting from a collection of manually labelled examples, or automatically, by means of a similarity measure that selects which texts have to be classified together and determines the final number of arbitrary domain categories (a classic clustering method). After the basic model generation (D_{nm}), a merging process, based on a percentage of resemblance (for instance 50%), was developed to obtain a hierarchical clustering of domains (D_n in Figure 1). The top of this structure is the full database, as in unit selection GP-TTS systems.

4 Experiments and evaluation

The proposed method was evaluated with a pair of experiments involving two collections of preclassified texts. C_{Cat} is a compilation of Catalan texts organized in 4 domains (*styles*): politics, society, literature and culture. C_{Cat} is an extension of the work described in [8]. C_{Spa} has been provided by the Centre de Lingüística Computacional (CLiC) from the Universitat de Barcelona (UB). It contains Spanish texts classified in 8 domains: the four included in C_{Cat} plus sports, business, entertainment and philosophy. The following tests were developed with relational graphs at basic model level (D_{nm}).

4.1 Experiment 1 - Text classification task

The TC algorithm was run over the C_{Spa} corpus (5000 sentences). Table 1 presents a summary of the TC micro-averaging precision (P^{μ}) effectiveness test [10]. This test is used to choose the best similarity measure for the MD-TTS system. The performance of the proposed algorithm for classic TC is not spectacular, however this is not its final purpose.

P^{μ}	4 domains		6 domains		8 domains	
%Test	**15**	**20**	**15**	**20**	**15**	**20**
S_1	.583	.471	.368	.343	.324	.322
S_2	.648	.635	.547	.566	.525	.529
S_3	.640	.628	.565	.533	.536	.520

Table 1: Micro-averaging precision of the three proposed similarity measures used for the TC algorithm over six different test configurations.

As a result, the inclusion of pl in S_2 enhances the cosine measure efficiency. S_3 (with $\mu = \overline{par}/2$ and $\sigma = \overline{par}/2$ in D_{nm}) is also a good distance metric, achieving similar results to S_2. Its efficacy improves when increasing the number of domains considered and the percentage of the training set. This is due to the tuning of μ and σ as the average for every *par* over all domains. Thus, adjusting them for every test configuration will certainly improve the performance of the algorithm.

4.2 Experiment 2 - MD-TTS synthesis task

The TC algorithm has been integrated into the MD-TTS system modelling the C_{Cat} corpus of domains (10000 sentences). This test evaluates the relationship between the Average Execution Time (AET) and the Average Segment Length (ASL) [8, 9] of the synthetic speech at sentence level. The time measures have been carried out over a Windows PC (PIV 1.6GHz - 256M RAM) using the Visual C++ compiler.

In the MD-TTS method, AET is the addition of the TC and unit selection temporal costs (let unit be representing diphones and triphones [8]). Moreover, the ASL reflects the potential speech quality of the synthesized sentence [9].

ISBN 1-876346-48-5

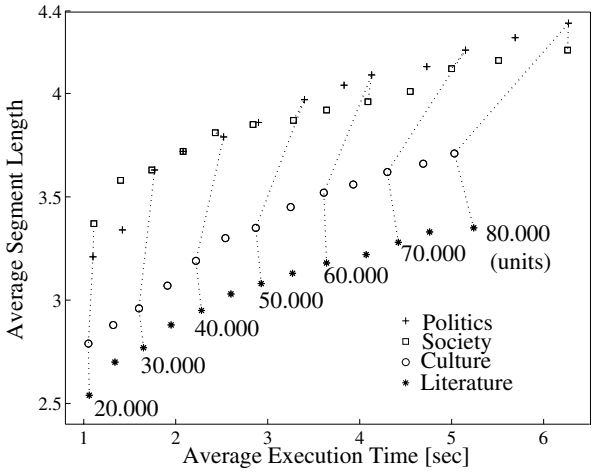

Figure 3: ASL *vs.* AET in four different Catalan domains and several D_{nm} dimensions (from 20k to 80k units, some of them interlaced by dashed lines).

Figure 3 plots the algorithm behavior found for various MD database sizes (in 5k units increments). The obtained curves are *log-like* functions. Politics and society tests had the better performance in ASL, since their higher mean unit length. However, the culture and literature tests had better AET performance regardless of database size (see the dashed lines).

Furthermore, the experiment has been developed for comparing the results obtained with a multi-domain database (M) against a full-domain database (F) (in table 2). The test is realized with 20k units per domain, thus, the F contains 80k units. This database is constructed as an ASL reference for the M database, but it is not a classic GP database. The test presents relative average reductions of 15% in ASL and 40% in AET when synthesizing in M *vs* F databases. Therefore, the computational cost of the unit selection process is reduced without a heavy loss of quality.

Test	**Politics**		**Society**		**Culture**		**Literat.**	
DB	F	M	F	M	F	M	F	M
ASL	3.2	2.6	3.5	2.8	4.0	3.3	4.0	3.4
AET	1.5	1.1	1.6	1.0	2.3	1.1	2.2	1.1

Table 2: Testing ASL *vs* AET [sec] in four domain (M) database (each of 20k units) against the full (F) database (DB) derived from it (80k units).

5 Conclusions

This paper has presented a first step in our on-going research towards a MD-TTS system to cover the niche between GP-TTS and LD-TTS applications. Moreover, we have introduced a domain classification algorithm based on ARN. In order to adjust the TC algorithm to the TTS purpose, some distance measures have been examined. The MD-TTS approach decreases the cost of searching a full domain database and yet yields good speech quality. In future experiments we will repeat this experiment against a GP-TTS database. We believe that the speech quality will be increased too, since the synthesis is done in the most appropriate domain (style) of the database.

Acknowledgments

We would like to thank the Generalitat de Catalunya and the D.U.R.S.I. for their support under grant number 2000FI-00679.

References

[1] A.W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *EuroSpeech*, Rodes, 1997, pp. 601–604.

[2] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Joint Meeting of ASA, EAA and DAGA2*, Berlin, 1999, pp. 18–24.

[3] G. Coorman, J. Fackrell, P. Rutten, and B. Van Coile, "Segment selection in the L&H RealSpeak laboratory TTS system," in *ICSLP*, Beijing, 2000, vol. 2, pp. 395–398.

[4] A.W. Black, "Perfect Synthesis for all of the people all of the time," in *IEEE TTS Workshop 2002 (Keynote)*, Santa Monica, 2002.

[5] B. Möbius, "Corpus-based speech synthesis: methods and challenges," in *AIMS*, vol. 6.

[6] A.W. Black and K. Lenzo, "Limited Domain Synthesis," in *ICSLP*, Beijing, 2000.

[7] B. Möbius, "Rare events and closed domains: two delicate concepts in Speech Synthesis," in *4th ISCA Workshop on Speech Synthesis*, 2001, pp. 41–46.

[8] R. Guaus and I. Iriondo, "Diphone-Based Unit Selection for Catalan TTS Synthesis," in *Proceedings of TSD*, Brno, 2000, Springer.

[9] M. Chu, C. Li, P. Hu, and E. Chang, "Domain adaption for TTS Systems," in *ICASSP*, 2002.

[10] F. Sebastiani, "Machine learning in automated text categorisation," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[11] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, 1989.

[12] E. Rennison, "Galaxy of News: An Approach to Visualizing and Understanding Expansive News Landscapes," in *ACM Symposium on User Interface Software and Technology*, 1994, pp. 3–12.