Multi-domain Text-to-Speech Synthesis by Automatic Text Classification

Francesc Alías, Joan Claudi Socoró, Xavier Sevillano, Ignasi Iriondo and Xavier Gonzalvo

Dep. of Communications and Signal Theory. Enginyeria i Arquitectura La Salle
Ramon Llull University, Barcelona, Spain
{falias,jclaudi,xavis,iriondo,gonzalvo}@salle.url.edu

Abstract

This paper describes a multi-domain text-to-speech (MD-TTS) synthesis strategy for generating speech among different domains and so increasing the flexibility of high quality TTS systems. To that effect, the MD-TTS introduces a flexible TTS architecture that includes an automatic domain classification module, which allows MD-TTS systems to be implemented by different synthesis strategies and speech corpus typologies. In this work, the performance of a corpus-based MD-TTS system is subjectively validated by means of several perceptual tests.

Index Terms: text-to-speech synthesis, general purpose, limited domain, multi-domain TTS, speech corpus, domain classification, automatic text classification.

1. Introduction

The final purpose of any Text-to-Speech (TTS) system is the generation of *perfectly* natural synthetic speech from *any* input text. To that effect, the capacity of processing unconstrained text has historically taken priority over the naturalness of the message, i.e. striving the flexibility of the application at the expense of the achieved synthetic quality [1, 2] (aka general purpose TTS - GP-TTS- systems) (see *Approach A* in figure 1). In the course of time, an opposite approach has prioritized the development of high quality TTS systems by restricting the scope of the input text, i.e. reducing the difficulty of the task [1] (*Approach B* in figure 1), giving rise to the so-called limited domain TTS (LD-TTS) synthesis. Following this second approach, we introduced multi-domain TTS (MD-TTS) synthesis in order to synthesize among different domains with high synthetic speech quality [3]. By one hand, the TTS task difficulty is increased due to the management of multiple domains, and, by the other hand, the achieved speech quality is equivalent to that of LD-TTS when the input text is assigned to the correct domain. Thus, this approach can represent an advance towards perfect unconstrained speech synthesis (see figure 1). The MD-TTS strategy can constitute an added value factor in any human-computer interaction (HCI) system with different domains of communication, e.g. multi-domain dialog systems or multimodal systems, such as talking heads, among others.

This paper firstly relates the MD-TTS approach to multi-domain spoken language systems and previous multi-domain speech corpus typologies. Secondly, the architecture and the main contributions of the proposal are described. Then, the performance of the MD-TTS system is evaluated in several perceptual tests conducted on a corpus-based MD-TTS system that incorporates a multi-domain Spanish speech corpus and the automatic domain classification module described in [3]. Finally, we discuss several issues related to the proposal, outlining future research directions.

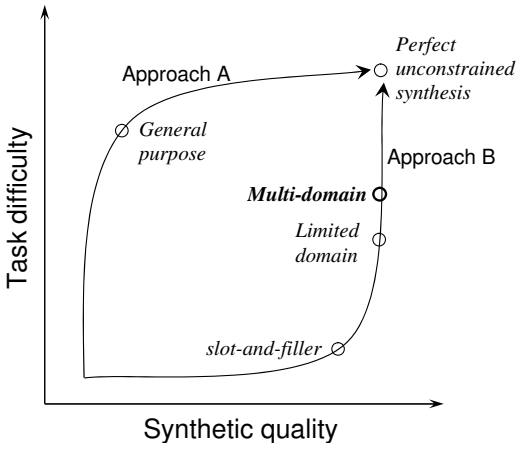


Figure 1: Different approaches to TTS research towards *perfect* unconstrained speech synthesis, as a function of the task difficulty and the obtained synthetic speech quality —adapted from [1, 2].

2. Related work

2.1. Multi-domain spoken language systems

The development of multi-domain applications is one of the new research directions in spoken language systems (SLS) [4]. Most of the SLS, excluding general purpose dictation systems, operate over a finite set of domains of interaction, e.g. different destinations in call routing, several topics in translation systems, or different subdomains in complex dialog systems [5]. Knowing the domain of communication —in this context, a *domain* generally corresponds to a *topic*—, allows to improve the performance and efficiency of the different SLS modules [6]. For instance, by selecting the most appropriate language model of a speech recognizer, by adapting the dialog manager strategy towards reducing the number of dialogue turns, or by dynamically loading the required resources according to the current domain of interaction [4, 5, 7].

Assigning domains to user utterances automatically is a key issue for multi-domain SLS (MD-SLS) [4]. The domain selection process can be *user-guided*, by explicitly using a predefined set of keywords, or *dialog-guided*, i.e. implicitly detected from speech recognition hypotheses [4, 5, 7]. The former simplifies the task of assigning domains but reduces the usability of the SLS. The latter allows natural navigation thanks to automatic topic classification, but has to deal with obtaining information from *short* utterances despite speech recognition *errors* [7] —a more complicated task than topic classification of articles or broadcast news [7, 8].

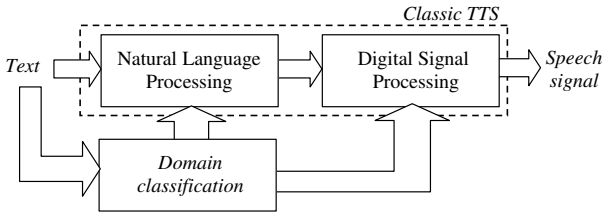


Figure 2: Block diagram of the multi-domain text-to-speech synthesis system including automatic domain classification.

2.2. Multi-domain speech corpora

A multi-domain speech corpus can be implemented in terms of different corpus typologies. There exist two main approaches from the corpus-based TTS point of view [9, 10], called: *i) tiering*, that is, defining an independent subcorpus for each domain [11, 12, 13, 14], and *ii) blending*, which consists in mixing different corpus subsets into an unique corpus, generally including a general purpose core [15, 16, 17]. Moreover, this issue has also been tackled by the HMM-based TTS research community (see [18] and related work), where these approaches have been called *style dependent modelling* and *style mixed modelling* —*style* corresponds to a particular kind of domain in MD-TTS context, where the term *domain* can stand for emotion, speaking style, topic, etc.

3. Multi-domain TTS synthesis

The quality of corpus-based TTS systems reflects very heavily the style and coverage of the recorded speech corpus [9, 10], decreasing when the input text mismatches the corpus domain coverage, for both GP-TTS [15, 17] and, more obviously, LD-TTS [19, 20]. In a previous work, the MD-TTS approach was defined as a first attempt to cover the niche between these approaches [3]. This technique is based on the fact that knowing the most appropriate domain for the input text allows much more proper delivery [9] —provided that that domain is properly synthesized from the speech corpus. There are sentences the meaning of which implies a specific style of delivery (e.g. positive or negative messages by using a particular prosodic pattern [16, 21]), while others may be unsuitable for certain speaking styles [18] (e.g. command utterances do not convey sadness or fear [12]). However, there are also messages the meaning of which depends on the context of communication [13], thus, the most appropriate speaking style depends on paralinguistic and extralinguistic information [22] —however, the study of this issue lies beyond the scope of this work.

In order to be able to infer the domain from the input text solely, it is necessary to redefine the classic TTS synthesis architecture by including a new module for conducting domain classification (see figure 2), as described in the following paragraphs.

3.1. MD-TTS system architecture

On the way towards improving the naturalness and usability of HCI systems, MD-TTS synthesis follows a counterpart evolution to MD-SLS. To that effect, the MD-TTS redefines the classic architecture of TTS systems by including a domain classification module, which interacts with both classic TTS modules, i.e. natural language processing (NLP) and digital signal processing (DSP) modules (see figure 2). Hence, in this context, knowing the domain of the input text allows to: *i)* help in the normalization process (e.g. if the input text belongs to a *mathematical* domain, the text "*1/2*" should be translated into "*half*" instead of "*January the second*"); *ii)* choose the most appropriate prosodic model for synthesizing that domain; *iii)* select the corresponding subcorpus for tiering approaches [12, 20] or guide the unit selection process by weighting accordingly the domain units for blending methods [16, 23]; *iv)* control the DSP module as regards the speech characteristics of that domain (e.g. vocal quality [24]); or *v)* activate the voice transformation module to resemble the target domain, if necessary; etc.

Furthermore, the MD-TTS architecture allows a flexible and adaptable TTS design and implementation that can be tuned according to the application needs or domain characteristics. Hence, the MD-TTS architecture can be adapted to whatever the most appropriate synthesis strategy (e.g. corpus-based, HMM-based, or hybrid solutions) and corpus typology (tiering or blending) are.

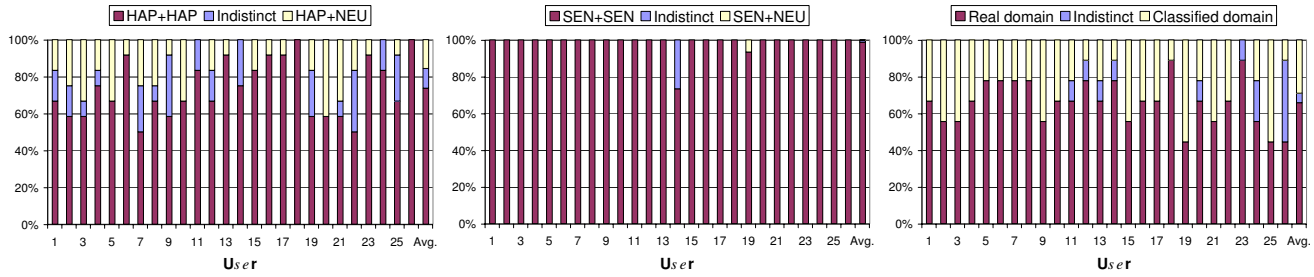
3.2. Domain classification

Storing different domains in a corpus is not a problem in itself if the TTS system is capable of managing this information appropriately [25]. As stated before, the MD-TTS system needs to know, during the TTS conversion process, which domain is the most suitable to conduct the synthesis so as to obtain the highest possible synthetic speech quality [9] and/or reduce the computational cost of the unit selection process [3]. The domain selection can be defined as an external task to the TTS system —manual [11, 13] or supervised selection— or it can be included as an automatic classification process. The supervised approach can be tackled e.g. by concept-to-speech synthesis [2] or by tagging the text [12, 20].

However, in order to include this task in the MD-TTS system, it is necessary to incorporate an automatic domain classification module in the classic MD-TTS architecture (see figure 2), hence, going further the typical text analysis of TTS systems (i.e. classic NLP capabilities). In the current version of our approach, this module is implemented by an automatic text classification algorithm based on a vector space model representation of texts, which includes information about the frequency and collocation of words plus the linguistic structure of text [3], in contrast to classic topic classification methods based on the *bag-of-words* approach [8]. Moreover, in its current version, this module is trained using the text corresponding to the recorded speech corpus (see section 4).

4. Experiments

After validating the viability of MD-TTS synthesis in [3], a *tiering* multi-domain Spanish speech corpus (2.5h) has been recorded by a female professional speaker to evaluate the proposal perceptually. This corpus consists of 2590 sentences extracted from an advertising database, which are grouped into three different domains: education (916 sentences), technology (833 sentences) and cosmetics (841 sentences). Each domain has been recorded using a predefined speaking style: happy (HAP), neutral (NEU) and sensual (SEN), respectively, regarding the contents of each domain. Thanks to the correspondence between speaking styles and domain contents, the automatic text classification module is able to select the most appropriate speaking style from text. The algorithm is trained on the 80% of corpus sentences and tested following a 10-fold random subsampling strategy. The current unit selection module is adjusted to extract the set of longest units from the classified domain, using a simple cost function [20]. The target prosody (pitch, duration and energy) is extracted from the real prosody of the tested sentence (*copy-prosody* strategy). However, if the classified domain is wrong, the prosody is predicted by the NLP module.



(a) Happy prosody + happy domain *vs.* happy prosody + neutral domain syntheses.

(b) Sensual prosody + sensual domain *vs.* sensual prosody + neutral domain syntheses.

(c) Real prosody + real domain *vs.* classified prosody + classified domain syntheses.

Figure 3: Preference tests between synthetic results from (a)(b) correct domain and (c) wrong domain classifications.

The following experiments are devoted to analyze the influence of correct and wrong domain classification decisions on the synthetic speech quality obtained by the MD-TTS approach when classifying texts as short as one sentence. It is to note that the average objective performance of the automatic text classifier across domains is, in this case, $F_1 = 0.78$, as the harmonic mean of precision and recall [8] —a forthcoming paper will detail this result. In the subjective experiments, the evaluators (26 members of our Dept.) were asked to select the most appropriate version between two synthetic results obtained from the same sentence, by means of a preference test including the *indistinct* option. The evaluators were able to listen to the generated files as many times as needed.

4.1. Subjective evaluation of correct domain classifications

The first test analyzes the achieved results when the correct domain of the input text is chosen. The results obtained by synthesizing in the correct domain are compared to the results attained from the neutral domain (used as reference regarding to what could be achieved by GP synthesis), both using the real prosodic pattern of the tested sentence. As a MD-TTS system which makes correct decisions is essentially a LD-TTS system in terms of speech quality, this experiment is equivalent to comparing LD-TTS to GP-TTS. The test was conducted on 12 sentences extracted from the happy domain and 15 sensual sentences, by applying a simple greedy algorithm tuned to select phonetically balanced sentences.

The results indicate a clear preference for the correctly classified domain outcomes over the reference syntheses, for both happy and sensual domains (see figures 3(a) and 3(b)). The leftmost figure shows an average preference greater than 80%, including indistinguishable selections. Moreover, no evaluator presented more than 40% of predilection on the general purpose based results. Furthermore, the test on the sensual domain reveals an overwhelming preference for the correct classification results compared to the reference. These results are mainly due to the peculiar characteristics of these domains (specially the *whispering* nature of the sensual domain), the synthesis of which, hitherto, has not been fully solved by prosodic modelling plus digital signal processing [24]. This is the main reason they have been explicitly contained in the speech corpus following a tiering approach, as in [11, 22].

4.2. Subjective evaluation of wrong domain classifications

The second test evaluates the perceptual impact of wrong automatic text classifications with respect to *a priori* labelling (i.e. *real* domain). Hence, this experiment is equivalent to comparing *worst-case* MD-TTS to LD-TTS synthesis. Each evaluator is asked to select the most appropriate synthetic version as regards the sentence meaning, since the style of delivery depends on the selected domain. As the tested sentences are assigned to a domain other than the one they have been originally recorded, the copy-prosody strategy is substituted by predicting the prosody from text by the NLP module (using one prosodic model per domain).

Figure 3(c) depicts the achieved results, obtained from 9 sentences misclassified by the automatic text classification module. As it can be observed, there is a slight preference for the real domain results rather than the incorrectly classified versions. However, the preference pattern is much less clear than the previous tests, exhibiting a higher deviation among users' elections, e.g. there are users with *real* / *classified+indistinct* ratios ranging from $8/1$ to $4/5$, but attaining a mean value of $6/3$. According to the users' perception, this experiment involved the most difficult elections (e.g. a larger number of turns were needed before deciding), as they were asked to take into account the message contents in their decisions instead of only comparing the synthetic speech qualities. Hence, the users showed different criteria when selecting the most appropriate delivery of the tested sentence, in contrast to the previous experiment, where a more homogeneous pattern was obtained (see figures 3(a) and 3(b)). The vague evaluators' criteria somehow correlate with the automatic domain misclassifications, which mostly occur due to sentences with no clear membership to any domain (e.g. "*The best solution*").

5. Discussion

The MD-TTS proposal is included in an incipient research direction towards incorporating deeper text analysis in the TTS systems so as to improve their synthetic speech quality. There are several recent papers focused on this issue by, e.g. extracting the user attitude from text [21] or guessing the underlying emotion of the message [23, 26] (see also references therein). Some of them are rule-based approaches (e.g. *Dictionary of Affect* in [23], or adjectives and adverbs lists in [21]), while others are based on machine learning techniques [26]. The MD-TTS system belongs to this second approach, however, its main characteristic is that, to date, it only takes into account the input text without including external knowledge, such as WordNet [26], although this possibility is left for future investigations to study if it can improve domain classification capabilities for MD-TTS synthesis.

Furthermore, the MD-TTS approach proposes a new direction towards improving the flexibility of high quality TTS systems, not just by optimizing the decomposition and reconstruction of the speech signal or explicitly including different domains in the corpus, as stated in [10], but also by allowing the inclusion of different synthesis strategies and corpus typologies in the TTS architecture through automatic text classification (a higher level of data organization). In this context, the MD-TTS system can be tuned to fit into the application requirements, instead of developing a TTS system from scratch for each new application (e.g. multiple LD-TTS systems working in parallel).

6. Conclusions

This paper has outlined our proposal towards improving the flexibility of TTS systems by considering multiple domains in the speech corpus and conducting automatic domain selection at run time, denoted as MD-TTS synthesis. As a first step towards this challenging goal, the MD-TTS approach has been implemented following a tiering corpus-based TTS strategy (due to the diverse voice qualities of the considered domains). The collected subjective results reveal that, when MD-TTS works appropriately (i.e. it is equivalent to LD-TTS), users prefer MD-TTS synthesis over GP-TTS synthesis, like in [15, 16, 17]. In contrast, when MD-TTS assigns the input sentence to a domain other than the one it has been originally recorded (i.e. wrong domain classification), evaluators showed rather vague preference criteria between synthetic results, as two different LD syntheses are being compared in this case. However, it is to note that with the current implementation of the text classification module [3], there is still room for further research when classifying text as short as one sentence. Moreover, this algorithm is currently being optimized towards reducing computational cost and improving classification efficiency. The number of domains of the multi-domain corpus is also being increased in order to conduct new subjective experiments.

7. Acknowledgements

This work was partly supported by the IntegraTV-4all project (grant no. FIT-350301-2004-2) of the Spanish Science and Technology Council. The authors would like to thank all participants involved in the subjective experiments.

8. References

[1] J. Yi and J. Glass, "Natural-sounding speech synthesis using variable-length units," in *Proceedings of ICSLP*, Sydney, Australia, 1998, pp. 1167–1170.

[2] P. Taylor, "Concept-to-Speech synthesis by phonological structure matching," *Philosophical Transactions of the Royal Society, Series A.*, vol. 356, no. 1769, pp. 1403–1416, 2000.

[3] F. Alías, I. Iriondo, and P. Barnola, "Multi-domain text classification for unit selection Text-to-Speech Synthesis," in *Proc. of the 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, 2003, pp. 2341–2344.

[4] K. Rüggenmann and I. Gurevych, "Assigning domains to speech recognition hypotheses," in *Proc. of HLT-NAACL Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing*, Boston, USA, 2004, pp. 70–77.

[5] I.R. Lane, T. Kawahara, T. Matsui, and S. Nakamura, "Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching," *IEICE Transactions on Information and Systems*, vol. E88D, no. 3, pp. 446–454, 2005.

[6] J. Allan, "Perspectives on Information Retrieval and Speech," *Lecture Notes in Computer Science (Workshop on Information Retrieval Techniques for Speech Applications)*, vol. 2273, pp. 1 – 10, 2001.

[7] K. Asami, T. Takezawa, and G. Kikui, "Topic detection of an utterance for Speech Dialogue Processing," in *Proceedings of ICSLP*, Denver, USA, 2002, pp. 1977–1980.

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

[9] A.W. Black, "Perfect Synthesis for all of the people all of the time," in *IEEE Workshop on Speech Synthesis*, Santa Monica, USA, 2002.

[10] A.W. Black, "Unit Selection and Emotional Speech," in *Proc. of EuroSpeech*, Geneve, Switzerland, 2003, pp. 1649–1652.

[11] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "A Speech Synthesis System with Emotion for Assisting Communication," in *Proc. of the ISCA Workshop on Speech and Emotion*, Newcastle, North Ireland, 2000, pp. 167–172.

[12] W.L. Johnson, S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. LaBore, "Limited domain synthesis of expressive military speech for animated characters," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA, 2002, pp. 163 – 166.

[13] N. Campbell, "What type of inputs will we need for Expressive Speech Synthesis?," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA, 2002.

[14] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based Speech Synthesis System with Emotion," *Speech Communication*, vol. 40, no. 1,2, pp. 161–187, 2003.

[15] M. Chu, C. Li, P. Hu, and E. Cahng, "Domain adaptation for TTS systems," in *Proc. of ICASSP*, Orlando, USA, 2002, pp. 453–456.

[16] W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, and J. F. Pitrelli, "The IBM Expressive Speech Synthesis System," in *Proc. of ICSLP*, Jeju Island, South Korea, 2004, pp. 2577–2580.

[17] V. Fischer, J. Botella, and S. Kunzmann, "Domain Adaptation Methods in the IBM trainable Text-To-Speech System," in *Proc. of ICSLP*, Jeju Island, South Korea, 2004, pp. 1165–1168.

[18] J. Yamagasi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic Modelling of Speaking Styles and Emotional Expressions in HMM-based Speech Synthesis," *IEICE Transactions on Information and Systems*, vol. E88D, no. 3, pp. 502–509, 2005.

[19] W. Hamza and J.F. Pitrelli, "Combining the flexibility of speech synthesis with the naturalness of pre-recorded audio: a comparison of two approaches to phrase-splicing TTS," in *Proc. of InterSpeech*, Lisbon, Portugal, 2005, pp. 2585–2588.

[20] F. Alías, I. Iriondo, L. Formiga, X. Gonzalvo, C. Monzo, and X. Sevillano, "High quality Spanish restricted-domain TTS oriented to a weather forecast application," in *Proc. of InterSpeech*, Lisbon, Portugal, 2005, pp. 2573–2576.

[21] Y. Sagisaka, T. Yamashita, and Y. Kokenawa, "Generation and perception of F_0 markedness for communicative speech synthesis," *Speech Communication*, vol. 46, no. 1, pp. 376–384, 2005.

[22] N. Campbell, "Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech," *IEICE Transactions on Information and Systems*, vol. E88D, no. 3, pp. 376–383, 2005.

[23] G. Hofer, K. Richmond, and R.A.J. Clark, "Informed blending of databases for emotional speech synthesis," in *Proc. of InterSpeech*, Lisbon, Portugal, 2005, pp. 501–504.

[24] O. Turk, M. Schröder, B. Bozkurt, and L.M. Arslan, "Voice quality interpolation for emotional Text-to-Speech synthesis," in *Proc. of InterSpeech*, Lisbon, Portugal, 2005, pp. 797–800.

[25] A. Breen and P. Jackson, "Non-uniform unit selection and the similarity metric within BT's LAUREATE TTS system," in *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 201–206.

[26] C. Ovesdotter, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of HLT/EMNLP*, Vancouver, Canada, 2005, pp. 579–586.