Text Classification based on Associative Relational Networks for Multi-Domain Text-to-Speech Synthesis

Francesc Alías, Xavier Sevillano, Joan Claudi Socoró
Dept. of Communications and Signal Theory
Enginyeria i Arquitectura La Salle. Ramon Llull University
Passeig Bonanova, 8. Barcelona, Spain
{falias, xavis, jclaudi}@salle.url.edu

ABSTRACT

This work is a step further in our research towards developing a new strategy for high quality text-to-speech (TTS) synthesis among different domains. In this context, it is necessary to select the most appropriate domain for synthesizing the text input to the TTS system, task that can be solved including a text classifier (TC) in the classic TTS architecture. Since speech speaking style and prosody depend on the sequentiality and text structure of the message, the TC should consider not only thematic but also stylistic aspects of text. To this end, we introduce a new text modelling scheme based on an associative relational network, which represents texts as a weighted word-based graph. The conducted experiments validate the proposal in terms of both objective (text classification efficiency) and subjective (perceived synthetic speech quality) evaluation criteria.

1. INTRODUCTION

The final purpose of any Text-to-Speech (TTS) system is the generation of *perfectly* natural synthetic speech from *any* input text. In the quest for this goal, two complementary strategies, which constitute a trade-off between speech naturalness and system flexibility, have been followed [14, 15]: *i*) the general purpose TTS synthesis (GP-TTS), which strives the flexibility of the application at the expense of the achieved synthetic speech quality, and *ii*) limited domain TTS (LD-TTS), which prioritizes the development of high quality TTS systems by restricting the scope of the input text (e.g. a weather forecast application [3]). As an approach to improve the GP-TTS flexibility while maintaining a speech quality equivalent to that of LD-TTS, we introduced multi-domain TTS (MD-TTS) in order to synthesize among different domains with good speech naturalness [2].

To this end, the MD-TTS system needs to know, at run time, which domain is the most suitable for synthesizing the input text with the highest synthetic speech quality (e.g. in concatenative speech synthesis, this involves selecting the most appropriate speech units from the corpus). Thus, in order to develop an fully automatic MD-TTS system, it is necessary to go further than the typical text analysis of TTS systems (i.e. classic Natural Language Processing capabilities), redefining the classic architecture of TTS systems by including a text classification module for input text domain assignment (see figure 1).

Traditional text classification (TC) techniques, which are mainly focused on thematic categorization, employ document representations which just consider the occurrence of the terms that constitute the texts (often, after filtering function words and stemming), ignoring their relationships [10]. Despite topic information is useful for organizing the speech corpus contents, relying solely on thematic aspects of text is insufficient for considering the inherent sequential nature of speech (e.g. prosody, coarticulation, etc.). Thus, proper TC for MD-TTS should take into account both thematic and stylistic aspects of text. In order to model all this information, a novel text representation technique based on an associative relational network (ARN) [7] was introduced in [2]. In short, ARN-based text representations include weights for words and their co-occurrences plus information regarding the structure of text. Equally important, TC for MD-TTS must consider *all* the terms and punctuation marks appearing in the text, not only because function words filtering would induce the loss of valuable information concerning text structure, but also because texts input to TTS systems can be very short, e.g. only one sentence.

In this paper, we firstly describe the text representation based on ARN (see section 2). Secondly, several experiments regarding TC in the MD-TTS context are presented (section 3). Finally, the conclusions of this work are discussed in section 4.

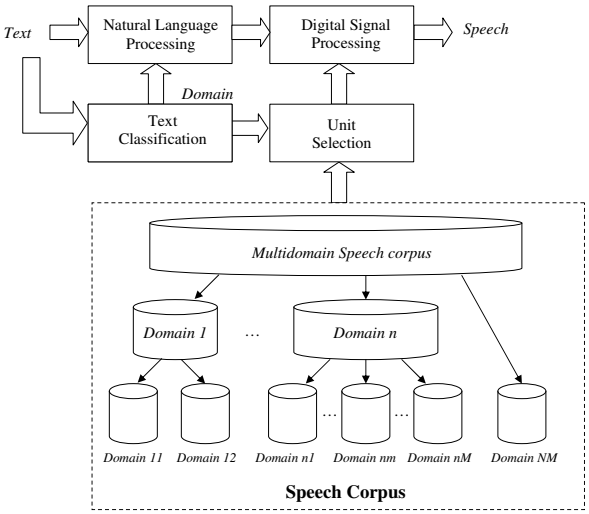

Figure 1: Block diagram of the MD-TTS system, according to the concatenative corpus-based TTS strategy.

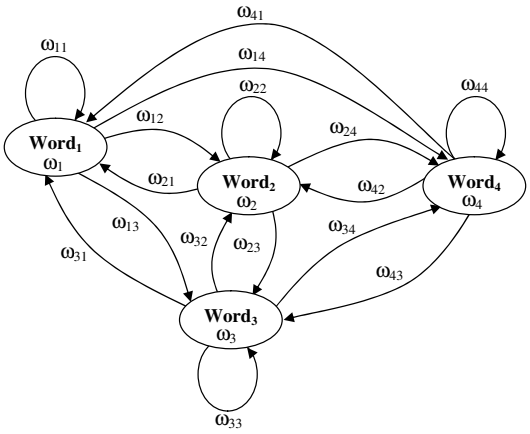

Figure 2: Word-based associative relational network, inspired by [7].

2. REPRESENTING TEXTS BY ASSOCIATIVE RELATIONAL NETWORKS

Most TC strategies treat texts as a collection of isolated words (the so called *bag-of-words* approach), i.e. each text is represented by its constituting terms, ignoring their order and relationships. However, there are several approaches that tackle non-thematic TC tasks, such as authorship attribution or genre detection [13]. In this context, function words distribution, part-of-speech tagging or word and sentence length, among others, are the most usually employed features (e.g. see [13]). However, these features are unable to represent fundamental stylistic factors synthetic speech quality depends upon, such as:

- **Sequentiality**: the synthetic speech signal is constituted by a sequence of consecutive speech units (phones, diphones, etc.) extracted from the speech corpus. If the text input to the TTS system was represented by means of isolated words, its inherent sequential nature would be lost (e.g. coarticulation effects between consecutive words would be missed).

- **Text structure**: the speech delivery pattern (i.e. speaking style and prosody: rhythm, tone, loudness, pauses, etc.) depends on the structure of text, which is embedded in word order and punctuation marks.

To deal with these two factors, it is essential to make use of a text representation technique capable of codifying both data. To this end, the developed TC system represents the texts by means of an Associative Relation Network (ARN), which was initially introduced in the context of visual representation of documents [7]. However, in our approach, the nodes of the graph represent the words of the text (including punctuation marks) and their connections describe the co-occurrences between words (see figure 2). Each node contains a weight (ω_i) assigned to its corresponding word and each connection is weighted by the relationship strength between the linked words (ω_{ij}), considering their order (not necessarily $\omega_{ij} = \omega_{j\,i}$).

As a result, the ARN is able to encode the sequentiality and the structure of the text, which are essential for classifying text in the MD-TTS framework.

2.1 Weighting the network

Once the ARN architecture is defined, it is necessary to assign specific values to the network weights. In particular, the nodes will contain thematic information while the internodal connections will be used to represent and extract stylistic patterns. For the time being, the thematic features (ω_i) employed are: *term frequency* $\times$ *inverse document frequency* (TFIDF) and a newly derived feature that we have called *inverse word frequency* (IWF), which is defined as:

$$iwf_i = log\left(\frac{M}{tf_i}\right), \forall\, tf_i > 0 \qquad (1)$$

where M is the number of words of the text and tf_i is the term frequency of the i-th term. IWF can be interpreted as a local approximation of IDF, since it weighs each term according to its prominence within *each* text, instead of considering its distribution across the *whole* text collection.

By its own definition, the ARN contains the *co-occurrence frequency* (COF) of each consecutive pair of words as stylistic information (ω_{ij}). However, this network architecture also allows considering structural resemblance between texts when conducting classification (see section 2.2.2).

2.2 Using ARN for text classification

In order to make use of the information embedded in the ARN, it is necessary to define a model suitable for conducting the classification task on a set of categories $\mathcal{C}$. To that effect, the TC system included in the MD-TTS architecture represents the ARN contents on a N-dimensional vector space model (VSM) [9], the dimensions of which correspond to the thematic and stylistic features extracted from text.

2.2.1 Training the ARN-based TC system

The training process consists in building an ARN for each of the $|\mathcal{C}|$ domains contained in the corpus from training documents $d_k \in \mathcal{D}$. In order to create a common representation space for all the domains, a *global* ARN is firstly built from all the training texts (see figure $3(a)$). Next, this global ARN is used as a reference for building each domain's ARN, obtaining what we have called Full ARN (ARN-F D_n, $n = 1 \ldots |\mathcal{C}|$), as its components follow the order indicated by the global ARN (see figure $3(b)$). The training stage finishes after deriving a vectorial representation of each ARN-F D_n yielding $|\mathcal{C}|$ pattern vectors ($\vec{p_n} \in \mathbb{R}^N$) within the VSM defined by the global ARN.

2.2.2 Classifiying the MD-TTS input texts

Given a text $t_k \notin \mathcal{D}$ input to the TTS system, it is firstly represented according to the global ARN model derived in the training stage, obtaining its corresponding vector $\vec{t_k} \in \mathbb{R}^N$. Hence, this vector can be compared to each of the $|\mathcal{C}|$ pattern vectors by simply computing a cosine similarity distance [9]. Nevertheless, the classification can be enriched with stylistic information by including a multiplicative factor that we call *pattern length* (PL) [2]. PL is defined as the length of the longest sequence of identical consecutive words appearing in the same order in both compared ARNs, normalized by the total number of words, thus, $0 \leq \text{PL} \leq 1$.

In this work, we introduce a subtle variation of PL, named cumulative PL (cPL), which is defined as the sum of consecutive co-occurrences matching between the input text and each pattern vector (see equation (2)).

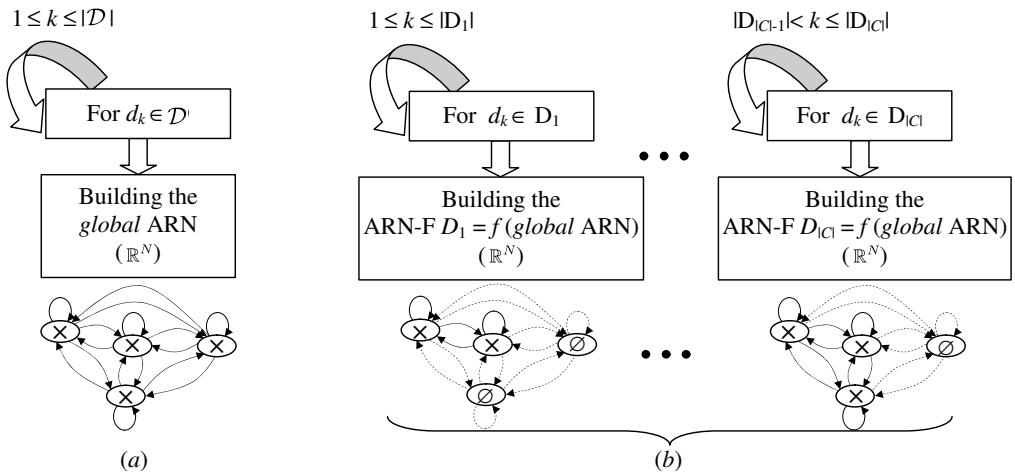

Figure 3: Building the (*a*) *global* ARN and (*b*) domain ARN-Fs, from D_1 to $D_{|\mathcal{C}|}$, according to the *global* ARN representation ($\mathbb{R}^N$) built from their corresponding training documents $\mathcal{D} = \{D_1, \ldots, D_{|\mathcal{C}|}\}$. In the graphs, "×" denotes filled nodes, "∅" represents empty nodes and dashed lines symbolize inexistent co-occurrences.

$$\mathrm{cPL}(\vec{t}_k, \vec{p}_n) = \frac{\sum_{i,j=1}^{M} \omega_{ij}^{\vec{t}_k}}{M-1} \tag{2}$$

Finally, the input text is assigned to that domain attaining the highest similarity in terms of a (stylistically weighted) cosine distance —see equations (3) to (5).

$$S_1(t_k, D_n) = \frac{< \vec{t}_k, \vec{p_n} >}{\|\vec{t}_k\| \cdot \|\vec{p_n}\|} \tag{3}$$

where $< \vec{a}, \vec{b} >$ denotes scalar product between vectors $\vec{a}$ and $\vec{b}$ and $||\vec{a}||$ represents the norm of vector $\vec{a}$.

$$S_2(t_k, D_n) = \mathrm{PL}(\vec{t}_k, \vec{p}_n) \cdot S_1(t_k, D_n) \tag{4}$$

$$S_3(t_k, D_n) = \mathrm{cPL}(\vec{t}_k, \vec{p}_n) \cdot S_1(t_k, D_n) \tag{5}$$

2.2.3 Reduced ARN model

Due to the extremely high dimensionality of the common representation space (full text representation and co-occurrence inclusion), the classification of each input text t_k is a computationally demanding task, since it requires going through the whole global ARN before conducting classification. Moreover, the vectorial representation of t_k will be typically *very* sparse, which results in a reduction of the separability properties of the pattern vectors, yielding poorer text classification efficiency. In order to improve domain separability and minimize the computational cost of the classification task, a novel ARN-based strategy called Reduced ARN (ARN-R) is introduced.

The main idea of the ARN-R model is based on the substitution of the full comparison space (built from the *global* ARN) by the VSM built from the ARN generated from the input text t_k. Hence, during the classification stage, each domain will be represented according to the ARN-R before conducting the comparison. That is, the domain ARN building process depicted in figure 3 is now conducted by substituting the global ARN by the ARN generated from the input text t_k. In this sense, the computational complexity of representing t_k on the global ARN space is substituted by the cost of representing each domain in the ARN-R space, which in general will be much lower.

Clearly, the ARN-R is just an approximation of the complete data representation provided by the ARN-F, as the ARN-R misses most of the information stored in the full space extracted from the training documents $\mathcal{D}$. Anyhow, it can be algebraically proved that the ARN-R is the best possible approximation of the ARN-F in the least mean square sense —further details can be found in [1].

3. EXPERIMENTS

The experiments have been conducted on a speech corpus composed of 1367 sentences extracted from an advertising database, which are grouped into three different domains: education (527 sentences), technology (323 sentences) and cosmetics (517 sentences). The studied TC algorithms are trained on the 80% of corpus sentences and tested following a 10-fold random subsampling strategy. In order to evaluate the performance of the TC algorithms in terms of classification efficiency (F1 measure [10]), the sentences have been randomly grouped into *pseudo*-doc-uments (hereafter, documents). This allows to analyze the performance of the TC methods as the number of sentences per document decreases, moving from a *standard* TC task (with many sentences per document) to a *typical* TTS scenario, with only one sentence per document. To that effect, a sweep ranging from 30 to 1 sentence per document is conducted.

3.1 Baseline method

The goal of this first experiment is to select a baseline TC algorithm to be used as a reference to validate the performance of the ARN-based TC proposals.

Specifically, three completely different TC strategies are compared, covering different approaches for solving the problem. Firstly, a basic Nearest Neighbour (NN) classifier [10] using TFIDF weighted terms as features is analyzed. This technique is based on representing each document as a vector in a VSM built from the training set. At classification time, each test document is assigned to the category of the most similar training document, according to a cosine distance. Secondly, a probabilistic TC algorithm based on *bigrams*

is also analyzed. In this case, each domain is represented by its own probabilistic language model obtained from the character pairs distribution across the documents of that domain [5]. The input text is assigned to the domain attaining the highest membership probability. Finally, an independent component analysis (ICA) based TC is applied to the problem. This technique, based on considering topics as latent random variables, makes use of term extraction for better thematic identification. In previous works, the ICA-based TC has been successfully applied for semi-supervised text classification and hierarchization of document corpora, by identifying the correspondence between text independent components and domains [11].

Figure 4 depicts the performance of the compared methods in terms of F1 across the sweep. It can be observed that the NN method shows the best global behaviour, followed by the probabilistic TC, whereas ICA-based TC suffers a rapid worsening due to the fact that this is a predominantly thematic approach. Hence, the NN classifier is selected as the baseline for validating the ARN-based proposals.

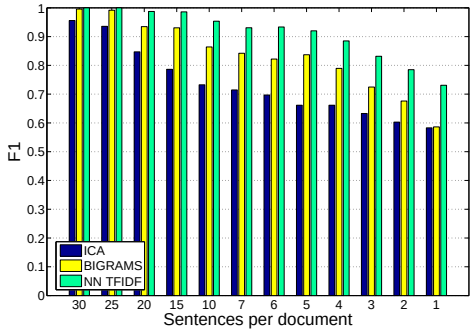

Figure 4: Classification efficiency of the baseline methods across the sentences per document sweep.

3.2 Performance of the ARN-based proposals

After selecting NN as the baseline method, the following paragraphs are devoted to validating the performance of the proposed ARN-based classifiers (ARN-F and ARN-R). To that effect, four different text parameterizations are considered. We compare the influence of using TFIDF *vs.* IWF as thematic features, besides considering stylistic information by means of COF or not (NCOF) in the ARN. On the other hand, we also analyze the impact of using similarity measures which incorporate stylistic information by means of PL and cPL.

3.2.1 *Text parameterization*

Figure 5 presents classification efficiency results of the compared TC techniques across the studied sweep, using the cosine distance as the similarity measure. On one hand, it can be observed that the ARN-based methods achieve better results than the baseline classifier (NN). However, both global representation methods (ARN-F and NN) are negatively affected by the inclusion of COF, achieving their optimal performance for TFIDF NCOF parameterization. In contrast, ARN-R even experiences a slight performance improvement when COF is considered. Moreover, ARN-R achieves its optimal performance when IWF is selected as the thematic feature. As a conclusion, it can be stated that ARN-R, despite being an approximation of ARN-F, behaves

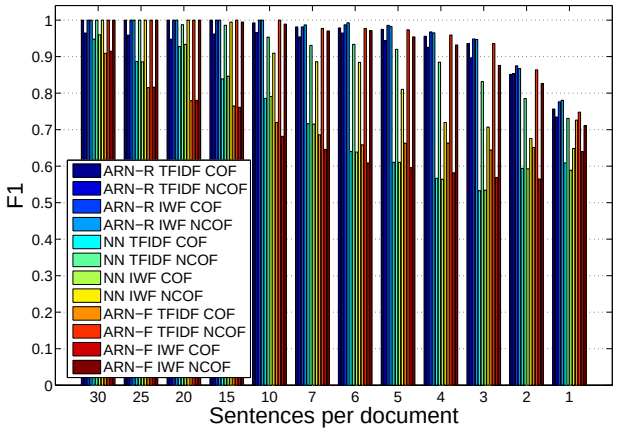

Figure 5: Classification efficiency of the ARN-based and NN TC methods across the sentences per document sweep for different text parameterizations.

more robustly in terms of the parameterization employed besides achieving equal or slightly better classification results in every step of the sweep (in particular, ARN-R is the best classifier in the hardest categorization scenario, i.e. 1 sentence per document).

3.2.2 *Similarity measures*

Figure 6 presents a global comparison regarding the use of stylistically weighted similarity measures for both ARN-based TCs. It can be observed that ARN-F experiences a notable improvement when the cosine distance is enriched with PL and cPL, attaining an average relative improvement of 14.2% and 19% on F_1, respectively. On the contrary, ARN-R is nearly unaffected by the inclusion of these factors in the similarity measure. As a conclusion, the stylistic weighting of distance measures affects the ARN-F-based TC positively, whereas this effect is less clear for the ARN-R classifier. Nevertheless, we shall study more deeply these results for ARN-R in future investigations.

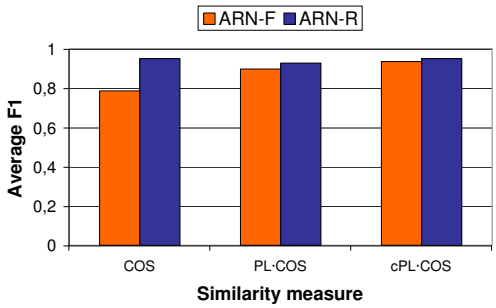

Figure 6: Averaged classification efficiency of the ARN-based TC methods across the sweep for different similarity measures.

3.3 Subjective results of the MD-TTS system

As the final goal of introducing a text classifier into the MD-TTS system architecture is to achieve high quality synthetic speech among different domains, a listening preference test was conducted by 26 evaluators in order to validate its

Table 1: Preference tests results for correct sentence classification.

Preference	Happy	Sensual
In-domain	74%	99%
Indistinct	11%	1%
Neutral	15%	0%

naturalness subjectively. Due to the fact that our MD-TTS approach is corpus-based, each domain has been recorded (in Spanish by a professional speaker) using a predefined speaking style regarding its contents: *happy* for education, *neutral* for technology and *sensual* for cosmetics.

The first subjective test puts in comparison the speech generated when the input text is assigned by the ARN-based TC to the correct domain and the speech synthesized from the neutral domain (used as a reference regarding to what could be achieved from GP corpus). The test was conducted on 12 sentences extracted from the happy domain and 15 sentences collected from the sensual domain. The results of this preference test (presented in table 1) denote an overwhelming preference for the correctly classified domain outcomes over the reference syntheses (achieved from the neutral domain), specially for the sensual domain —due to its particular whispering nature.

Subsequently, a second preference test was conducted so as to evaluate the perceptual impact of wrong text automatic text classifications with respect to *a priori* labelling (i.e. in-domain synthesis). Hence, this experiment is equivalent to comparing *worst-case* MD-TTS to LD-TTS synthesis. To that end, 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. In this case, the preference results show a less clear trend, i.e. the preference pattern among users showed a greater variation compared to the previous test, though there is a slight preference for the in-domain results: 66% vs. 34%, including indistinctness —see [4] for a more detailed analysis.

4. CONCLUSIONS

The MD-TTS proposal is included in an incipient research direction towards including deeper text analysis in 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 [8] or guessing the underlying emotion of the message [6] (see also references therein). The described ARN-based TC tackles satisfactorily the problem of classifying texts as short as one sentence, by taking into account both thematic and stylistic features (within the text representation and/or weighting the similarity measure). Moreover, the conducted subjective experiments show a nice correlation between evaluators' preferences and TC assignments, validating the performance of the ARN-based TC perceptually. However, there is still room for further research by, for instance, conducting feature ensembling for improving the classification efficiency —see [12] as a first attempt to this goal.

5. ACKNOWLEDGMENTS

This work was partly supported by the IntegraTV-4all project (grant no. FIT-350301-2004-2) of the Spanish Science and Technology Council.

6. REFERENCES

[1] F. Alías, X. Gonzalvo, X. Sevillano, J. Socoró, J. Montero, and D. García. Text classification adapted to Multi-domain Text-to-Speech Synthesis. *Procesamiento del Lenguaje Natural*, 37, September 2006 (*In Spanish*).

[2] 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)*, pages 2341–2344, Barcelona, Spain, 2003.

[3] 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*, pages 2573–2576, Lisbon, Portugal, 2005.

[4] F. Alías, J. Socoró, X. Sevillano, I. Iriondo, and X. Gonzalvo. Multi-domain Text-to-Speech Synthesis by Automatic Text Classification. In *Proc. of InterSpeech*, Pittsburg, USA, 2006.

[5] W. Cavnar and J. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.

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

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

[8] Y. Sagisaka, T. Yamashita, and Y. Kokenawa. Generation and perception of F_0 markedness for communicative speech synthesis. *Speech Communication*, 46(1):376–384, 2005.

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

[10] F. Sebastiani. Machine learning in automated text categorisation. *ACM Computing Surveys*, 34(1):1–47, 2002.

[11] X. Sevillano, F. Alías, and J. Socoró. ICA-Based Hierarchical Text Classification for Multi-domain Text-to-Speech Synthesis. In *Proceedings of ICASSP*, volume 5, pages 697–700, Montreal, Canada, 2004.

[12] X. Sevillano, G. Cobo, F. Alías, and J. Socoró. Feature Diversity in Cluster Ensembles for Robust Document Clustering. In *The 29th Annual International ACM SIGIR*, Seattle, USA, 2006.

[13] E. Stamatatos, G. Kokkinakis, and N. Fakotakis. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471 – 495, 2000.

[14] P. Taylor. Concept-to-Speech synthesis by phonological structure matching. *Philosophical Transactions of the Royal Society, Series A*, 356(1769):1403–1416, 2000.

[15] J. Yi and J. Glass. Natural-sounding speech synthesis using variable-length units. In *Proceedings of ICSLP*, pages 1167–1170, Sydney, Australia, 1998.