ISCA Archive
http://www.isca-speech.org/archive

2nd Workshop on Speech, Language
and Audio in Multimedia (SLAM 2014)
Penang, Malaysia
September 11-12, 2014

Development of a Genre-Dependent TTS System with Cross-Speaker Speaking-Style Transplantation

Jaime Lorenzo-Trueba[1], *Julián D. Echeverry-Correa*[1], *Roberto Barra-Chicote*[1], *Rubén San-Segundo*[1], *Javier Ferreiros*[1], *Ascensión Gallardo-Antolín*[4], *Junichi Yamagishi*[2,3], *Simon King*[2], *Juan M. Montero*[1],

[1] Speech Technology Group. E.T.S.I. Telecomunicación. Universidad Politecnica de Madrid (UPM)
[2] Centre for Speech Technology Research (CSTR), University of Edinburgh
[3] National Institute of Informatics (NII), Tokio
[4] Universidad Carlos III de Madrid (UC3M)

{jaime.lorenzo, jdec, barra, lapiz, jfl}@die.upm.es, gallardo@tsc.uc3m.es,
jyamagis@nii.ac.jp, Simon.King@ed.ac.uk, juancho@die.upm.es

Abstract

One of the biggest challenges in speech synthesis is the production of contextually-appropriate naturally sounding synthetic voices. This means that a Text-To-Speech system must be able to analyze a text beyond the sentence limits in order to select, or even modulate, the speaking style according to a broader context. Our current architecture is based on a two-step approach: text genre identification and speaking style synthesis according to the detected discourse genre. For the final implementation, a set of four genres and their corresponding speaking styles were considered: broadcast news, live sport commentaries, interviews and political speeches. In the final TTS evaluation, the four speaking styles were transplanted to the neutral voices of other speakers not included in the training database. When the transplanted styles were compared to the neutral voices, transplantation was significantly preferred and the similarity to the target speaker was as high as 78%.

Index Terms: speech synthesis, speaking style transplantation, automatic genre identification, Latent Semantic Analysis

1. Introduction

One of the deepest problems in current speech synthesis systems is the ubiquitous use of neutral speaking styles. They are designed to be used in any context, although in reality they are clearly inappropriate and unnatural in almost all real applications and listening situations. Therefore, conversational and expressive styles (as opposed to ‘read text’) are an important target in FP7 Simple4All project [1], as part of a suite of techniques and tools which can cost-effectively construct a speech synthesizer that fits a specific context of use.

Since one of the objectives in Simple4All is the development of flexible models capable of producing a range of expressive or conversational speaking styles, this paper will describe several genres and speaking styles from textual and prosodic points of view, plus the results of our first experiment on speaking style modeling and transplantation. Given the flexibility needed for transplanting styles, HMM-based synthesis is a natural choice: HMM-based synthesis, on the other hand, due to its parametric nature, it is much more flexible, a fact that can be exploited even further by using adaptation techniques [2]. Consequently this study focuses on HMM-based synthesis and adaptation techniques in order to produce voices with the desired speaking styles.

Although one could argue that each speaker has a specific style, a kind of vocal signature which makes the speaker identifiable, this personal speaking style is modulated according to context in order to adapt it to that context. Different kind of texts (sometimes called discourse genres) have different lexical, syntactical and semantic features which make the genre identifiable, but they are also linked to a certain speaking style which best suits that genre. When reading a text or playing a role, speakers are able to identify the genre of the text, and they are also able to adapt their speaking style to that genre (most of the time at least) by adopting conventions generally associated to that particular genre.

The main objectives of a project such as Simple4All (when dealing with speaking styles) would be the following ones:

- to select a set of discourse genres and a set of associated speaking styles which are textually and acoustically different and which are able to cover a broad range of speaking situation and speech applications,
- to automatically extract text features to be used to identify the genre of a given text and to train classifiers to predict the genre of that text,
- to train and evaluate the speaking style Text-To-Speech models which can be associated to the predicted discourse genre.

With the rapid growth of the information available online, Automatic Genre Identification (AGI) has become a key technique in text data classification [3]. This technique addresses the problem of identifying which genre best matches a certain textual document, given a limited predefined set of genres. It is currently being used in many domains of applications when document classification is needed, ranging from document indexing and automatic metadata generation, to message filtering. The task of AGI falls at the intersection of information retrieval and machine learning systems. In the last years a growing number of statistical learning methods have been applied in AGI [4].

A standard text identification framework comprises several steps: preprocessing, feature extraction, feature selection and classification. The preprocessing module usually contains several stages: tokenization, stop-word removal, stemming and, word categorization or Part-Of-Speech tagging.

Regarding the feature extraction module, the most usual approach is the vector space model [5], which is based on the use of a bag of words [6]. The feature selection module generally uses filtering methods (such as weighting schemes for the term frequency TF and the inverse document frequency IDF [7]), or techniques for computing the mutual information of terms [8], information gain [9] and chi-square statistical metrics [10]. The classification module can use well-established techniques from the fields of information retrieval and machine learning, such as Latent Semantic Analysis [11], Decision Trees [12], Naïve Bayes classifiers or Support Vector Machines [13].

Regarding modeling speaking styles, this paper goes beyond the standard approaches which create models for every style in the dataset (including one for neutral or read-speech voices). In addition to that standard modeling, the approach in this paper computes the differences between the neutral model and the target speaking style which are modeled through adaptation transformations. This way, it is possible to transferring these differences into a new target speaker in order to modify the recorded speaking style of that speaker and to adapt it to the genre of the input text. This cross-speaker transformation allows generating combinations of voices and speaking styles which have never been recorded, and to infuse expressiveness into non-expressive previously-recorded speaker models.

The outline of the papers is as follows: in section two the datasets used will be described; section two focuses on the experiments on genre classification using the two available dataset; section four describes the style transplantation scheme and the results of the perceptual test; finally sections five and six include the conclusions and future work.

2. Databases

2.1. C-ORAL-Rom

C-ORAL-Rom database is a multi-language multi-style database [14]. Out of all the available data, only four of the styles available in the Spanish formal media section have been used in this study: news, sports, scientific reports and interviews (a fifth class ‘others’ was dismissed after initial experiments; ‘others’ combined several styles such as reportage and talk shows). Among all the available data of each style, a subset of the least noisy audio was selected.

2.2. Spanish Speaking Styles (SSS)

Although there is another available Spanish speaking styles databases such as Glissando [17], neither C-ORAL-Rom nor Glissando were appropriate for the full set of experiments carried out for this paper, because of the number of styles was reduced or the recording quality was not high and because each speaker was recorded in only one speaking style without any overlapping. This kind of overlapping is necessary for developing the transplantation module.

Therefore, we have designed and recorded a new Spanish Speaking Styles database (SSS). We have recorded one male speaker in several speaking styles (including a reference read style). The main recorded styles this paper is focused on are: broadcast news, interviews, political speeches and live sport commentaries. We have also recorded several secondary styles for future use (child, witch, ogre and old man). For each main style, we have recorded about one hour. The prompted styles (news and political speeches) were recorded in paragraphs; improvised styles (interviews and live sports) were recorded continuously, and segmented offline. Three types of pauses were hand labeled: intra-sentence pauses, inter-sentence pauses and filled pauses. This labeling was imposed by difficulties when modeling interviews, due to its extremely high number of filled pauses and repetitions.

A prosodic analysis in Figure 1 shows the main speaking styles in SSS are quite separable, except for the conversational interviews, which is quite broad.

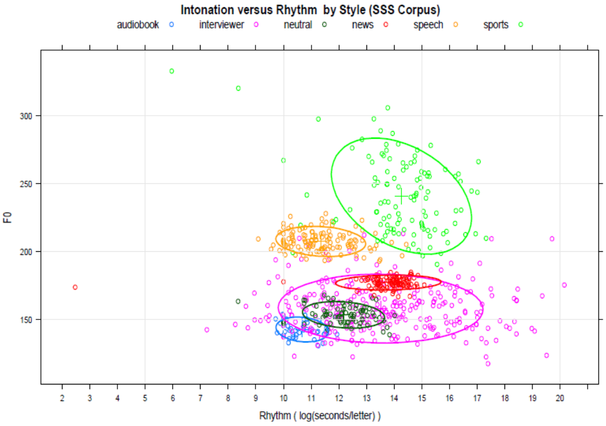

Figure 1: *Prosodic map of the main speaking styles in the SSS database*

3. Experiments on genre classification

3.1. C-ORAL-ROM experiments

The XML transcription files were pre-processed and divided into turns or sentences. As we were dealing with transcriptions from audio files, there were no punctuation marks or non standard words on the input text, and punctuation marks could not be used as input features. We tested the TF-IDF approach, with a minimum frequency threshold, a list of stop-words and a stemming module (Spanish Freeling [16]). The best results were obtained when filtering the data by removing the shortest turns (shorter than 5, 10 or 15 words). We evaluated several Weka classifiers (such as Multiclass classifier, Logistic, LogiBoost, Classification via regression, Simple logistic, LMT tree, Naïve Bayes, SVM-SMO, Random subspace) [19]. The best results were obtained when using the Naive Bayes Multinomial classifier.

There was no improvement over the TF-IDF results when using language model features such as: probability features (maximum and minimum), N-gram features (maximum, minimum, average and percentage of 1-gram), N-gram loss (maximum, minimum and average), Perplexity (PP), Out of vocabulary words contribution to PP, number of words found in vocabulary words, number of words out of vocabulary, percentage of out of vocabulary words…

The best results were obtained when using a minimum term frequency of 3, a list of 800 terms with the best TF-IDF value, removing turns with less than 15 words and not using the stemming module, achieving a 88.65% accuracy on the 4-class problem and 78.99% on the 5-class problem, both are significantly better than their baselines (80.84% and 83.87%, respectively) and very appropriate for their use in a TTS system.

Stemming did not contribute to improve the results. This may be caused because of the lost of semantic information when reducing words to their stems and thus the relationships between terms and documents may be distorted for this approximation.

3.2. Experiments on the SSS corpus

The new SSS database is much more homogeneous than C-ORAL-ROM, and much more appropriate for speaking style modeling and transplantation.

When carrying out experiments on genre prediction using the data from SSS and a Latent Semantic Analysis (LSA) approach (which decomposes the Term-Document Matrix using Singular Value Decomposition) and a cosine similarity function of the text to be classified and the centroids of the genres according on the training set.

LSA assumes that there is some underlying structure in word usage that is partially obscured by variability in word choice. To reveal this structure, LSA projects documents and queries into a space with latent semantic dimensions. In this space two texts can have high cosine similarity even if they do not share any terms. This is possible as long as their terms are conceptually similar or as long as they have been used to express similar concepts. The latent semantic space has fewer dimensions than the original space (which has as many dimensions as index terms) and can be more robust.

The achieved accuracy was as high as 94% in a 4-class problem, with the confusion matrix (on the evaluation set) shown in Table 1. These result were significantly higher than those obtained using the previous TF-IDF approach (Deliverable 5.2 at [1])

Table 1. *Confusion matrix of genre prediction in the SSS database.*

	Interview	News	Political speech	Sport comments
Interview	19/21	1/21	1/21	0
News	1/21	20/21	0	0
Political speech	0	1/21	20/21	0
Sport comments	0	0	1/21	20/21

4. Speaking-Style Transplantation experiments

Once we have been able to guess the genre of a certain text, we can assign an appropriate speaking style for that text. Given that not all the styles are initially available for all the speakers, it would be great to be able to transplant the speaking style (but not the speaker vocal identity) from a professional actor or actress to any speaker. This way all the available synthetic voices could benefit from the genre detector, and would be able to adapt the speaking style to each text without having acting skills or going through an acting training processes.

In a final perceptual test we have used the technique described in [15], to transplant the speaking styles into a male speaker (FFM) different from the SSS speaker (JOA), using an adaptation-based technique as shown in Figure 2. It was a preference test comparing a standard read style with the transplanted style: which is more appropriate for the text to be synthesized. The transplantation factor was 1.0.

The most significantly preferred styles (after transplantation) were sports (94% vs. 6%) and political speech (88% versus 12%). The news style is so similar to the standard read style, that it is not preferred after transplantation (38% vs. 69%), because any transplantation artifact can affect the synthesized quality.

Regarding interviews, the style is clearly preferred (56% vs. 44%) but the difference is only slight because it is a broad conversational style. Although some disfluencies have been hand labeled, the TTS system is not able to deal with those disfluencies. Most probably the amount of collected data is enough for the other more homogeneous styles, but not enough for coping with this very diverse style. In the future we should focus on a more specific domain and use domain-dependent semantic tags, and carry out some work on filled pause and disfluency modeling.

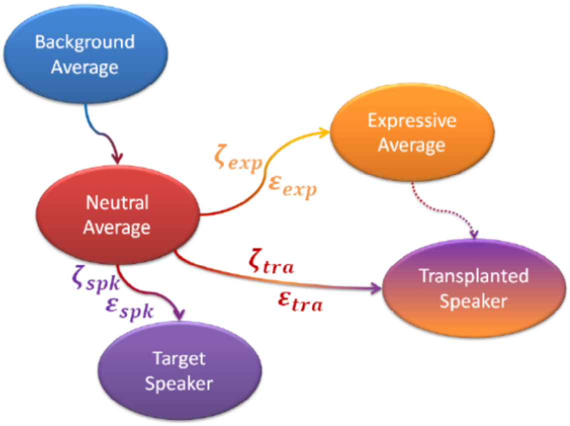

Figure 2: *Speaking style transplantation scheme: from an expressive source speaker into a style-transplanted target one.*

Regarding the similarity of the transplanted speech to the original speakers (JOA or the transplanted speaker), the average preference for the transplanted speaker over the source SSS speaker is high, about 78%, ranging from 63% for interviews to 100% for sports, and about 78% for the new transplanted speaker.

The overall MOS after transplantation is slightly lower than the standard speaker-dependent read style (2.54 vs. 2.38), but the strength of the transplanted synthesized style is significantly higher (2.48 vs. 1.82).

5. Conclusions

We have addressed the task of genre and sub-genre identification as a pre-processing tool for style-dependent TTS. Several criteria have been used for creating the stop-word list and therefore for selecting the index terms used in the identification task. Although stemming contributes to reducing the size of the indexing structure, it does not contribute to reducing the identification error. Language model features do not contribute to improve the identification rate.

In spite of the robustness of TF-IDF features, we have been able to get the best results by using entropy-based Latent Semantic Analysis classifiers, achieving a very high recognition rate (94%) on the SSS corpus used for TTS experiments on speaking style transplantation.

All the transplanted styles have been significantly preferred when compared to a neutral voice, and the similarity to the target speaker was as high as 78%, with small variation among speakers. The best transplanted styles were sport commentaries and political speeches.

6. Future work

In this paper, we have addressed the problem of synthesizing speaking styles using categorical labels (a set of discrete genres and speaking styles) and style-dependent synthesis models. However, by using the LSA technique, we can compute how similar a new text is when compared to the prototypical speaking styles in the training set and convert the problem into a semi-continuous interpolation problem, using the similarity function from the genre classifier to compute the transplantation weights to be applied to and reference average expressive style and to the detected one. If the projected vector of the text is very close to one of the genres, the transplantation will be close to the style corresponding to that genre; however, if the projected vector is far away from all the styles, the transplantation factor will be close to zero, and an average expressive style will be mostly used for synthesizing the text.

We could even use the LSA projected features as input features for improving the speaking-style synthesis training process, without any previous classification stage, developing a fully-continuous approach for expressive speaking style TTS.

7. Acknowledgements

Part of this work has been carried out during the research stay of A. Gallardo-Antolín and J. M. Montero at the Centre for Speech Technology Research (CSTR), University of Edinburgh, supported by the Spanish Ministry of Education, Culture and Sports under the National Program of Human Resources Mobility from the I+D+i 2008-2011 National Program, extended by agreement of the Council of Ministers in October 7th, 2011.

The work leading to these results has received funding from the European Union under grant agreement 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02) and TEC2011-26807 projects. Jaime Lorenzo-Trueba has been funded by Universidad Politecnica de Madrid under grant SBUPM-QTKTZHB.

Authors also thank the other members of the Speech Technology Group (especially Miguel Fernández, Jose Moreno and Arturo Romero) and the Simple4All project for the continuous and fruitful discussion on these topics, and Joaquin Martín (JOA) and Fernando Fernández (FFM), our speakers, for proving their voices for these experiments.

8. References

[1] Simple4All project http://simple4all.org

[2] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 1, pp. 66–83, 2009.

[3] P. Petrenz and B. Webber. Stable classification of text genres. Computational Linguistics, 37(2):385 – 393, June 2011.

[4] F. Sebastiani. Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1): pages 1–47, 2002.

[5] G. Salton, C.-S. Yang, and C. Yu. A theory of term importance in automatic text analysis. Journal of the American Society for Information Science, 26(1): pages 33–44, 1975.

[6] R. A. Baeza-Yates and B. A. Ribeiro-Neto. Modern Information Retrieval - the concepts and technology behind search, Second edition. Pearson Education Ltd., Harlow, England, 2011.

[7] S. T. Dumais. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments & Computers, 23(2): pages 229–236, 1991.

[8] H. Liu, J. Sun, L. Liu, and H. Zhang. Feature selection with dynamic mutual information. Pattern Recognition, 42(7): pages 1330 – 1339, 2009

[9] C. Lee and G. Geunbae-Lee. Information gain and divergence-based feature selection for machine learning-based text categorization. Information Processing and Management, 42(1): pages 155–165, January 2006.

[10] Y.-T. Chen and M. Chang-Chen. Using chi-square statistics to measure similarities for text categorization. Expert Systems with Applications, 38(4): pages 3085–3090, 2011.

[11] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 4, 41(6): pages 391–407, 1990.

[12] D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In Proceedings of the 1994 Symposium on Document Analysis and Information Retrieval, pages 81–93, 1994.

[13] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, pages 137–142, London, UK, UK, 1998. Springer-Verlag.

[14] A. Moreno-Sandoval, G. De la Madrid, M. Alcantara, A. Gonzalez, JM Guirao, and R. De la Torre. The Spanish corpus. C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages, Amsterdam: John Benjamins Publishing Company, pages 135–161, 2005

[15] J. Lorenzo-Trueba, R. Barra-Chicote, J. Yamagishi, O. Watts, and J. M. Montero. Towards speaking style transplantation in speech synthesis. In 8th ISCA Workshop on Speech Synthesis, pages 179–183, Barcelona, Spain, August 2013.

[16] L. Padró and E. Stanilovsky. Freeling 3.0: Towards Wider Multilinguality. In Proceedings of the Language Resources and Evaluation Conference, pages 2473 – 2479, Istanbul, Turkey, May 2012.

[17] J. M. Garrido, D. Escudero, L. Aguilar, V. Cardeñoso, E. Rodero, C. de-la-Mota, C. González, C. Vivaracho, S. Rustullet, O. Larrea, Y. Laplaza, F. Vizcaíno, E. Estebas, M. Cabrera, A. Bonafonte. Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. Language Resources & Evaluation 47(4): pages 945–971. 2013

[18] S. Brognaux, T. Drugman, M. Saerens, "Synthesizing sports commentaries: One or several emphatic stresses?", Speech Prosody, Dublin, Ireland, May 21-24, 2014.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1. 2009.