Towards an Unsupervised Speaking Style Voice Building Framework: multi–style speaker diarization

J. Lorenzo-Trueba[1], *B. Martinez-Gonzalez*[1], *V. Lopez–Ludeña*[1], *R. Barra-Chicote*[1], *J. Ferreiros*[1], *J. Yamagishi*[2], *J.M. Montero*[1]

[1]Speech Technology Group, ETSIT, Universidad Politecnica de Madrid, Spain
[2]Centre for Speech Technology Research, University of Edinburgh, United Kingdom
jaime.lorenzo@die.upm.es, barra@die.upm.es

Abstract

Current text–to–speech systems are developed using studio-recorded speech in a neutral style or based on acted emotions. However, the proliferation of media sharing sites would allow developing a new generation of speech–based systems which could cope with spontaneous and styled speech. This paper proposes an architecture to deal with realistic recordings and carries out some experiments on unsupervised speaker diarization. In order to maximize the speaker purity of the clusters while keeping a high speaker coverage, the paper evaluates the F–measure of a diarization module, achieving high scores (>85%) especially when the clusters are longer than 30 seconds, even for the more spontaneous and expressive styles (such as talk shows or sports).
Index Terms: expressive speech synthesis, speaker diarization, speaking styles, voice cloning.

1. Introduction

There has been a sharp increase in the use of computers in every aspect of daily life, from tablets to smart–phones, leading to a proliferation of media sharing sites which provide new data for developing speech–enabled applications. Under this prospects, and with the purpose of increasing accessibility and ease of use, it becomes very interesting to provide the machines with increasingly versatile human–machine interfaces, not only on speech recognition but also on speech synthesis. Unfortunately, current speech synthesis technologies show severe deficiencies when dealing with spontaneous human–like speech, a great obstacle when implementing widespread applications.

One of the goals in *The Simple4all Project* [1] is the automatic modification of the speaking style of a neutral or expressive voice without needing to record a new speaker under the target expressivity conditions, maintaining the quality of the original models but achieving a high style (or emotion) identification rate, and being able to control the intensity of the expressivity in a continuous range. The first step for being able to successfully control the speaking style of the synthetic speech is to obtain enough data from speakers with different speaking styles with which we could build speaking style average models.

Another goal of *Simple4all* is to create the most portable speech synthesis system possible: one that could be automatically (or with limited manual supervision) applied to many domains and tasks, which implies dealing with a wide variety of expressive situations and domains. In order to use speech collected from the media or from media sharing sites, speech synthesis systems must be robust to the variation of the acoustic and environmental conditions. The system mush be able to robustly cope with noisy ASR–processed corpora and with challenging data such as interviews, debates, home recordings, political speeches, etc. The use of diarization techniques for speaker–turn segmentation will allow the system creating homogeneous voices from heterogeneous recordings, because the number of speakers would be automatically estimated in a fully unsupervised way, and language–independent diarization techniques automatically could provide the temporal labels of the turns of a certain speaker [2, 3].

In this work we present the architecture of style–cloning system and an evaluation of the performance of a speaker diarization system for unsupervisedly–generating clusters of speech from several styled recordings. Those clusters will be used for building average style models in order to incorporate the expressiveness of those recordings to the synthetic voice of new speakers.

2. Unsupervised Multi-Speaker Expressive Voice Building Framework

Considering the amount of multimedia data currently available in the Internet and in the media (television, radio, pod-casts, audio–books, etc.) we define an unsupervised voice–building framework that will allow developing multiple voices with different speaking styles. The architecture shown in Figure 1 will minimise the manual

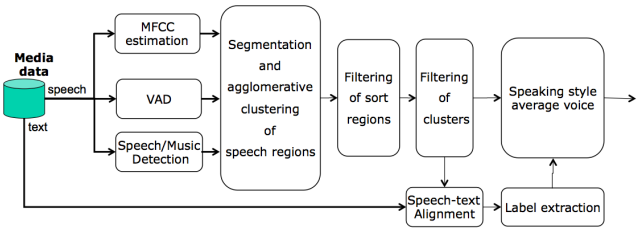

Figure 1: *Unsupervised Multi-Speaker Expressive Voice Building Framework*

processing of the multimedia data and the requirement of expert knowledge.

This framework imposes the use of many speech processing technologies. The implementation of this framework is reasonable due to the availability of open–source technologies: Voice Activity Detection (VAD) [4], Speaker Diarization [5], Speech/Music Segmentation [6], Automatic Speech Recognition [7, 8], automatic Speech-Text Time Alignment [9, 10], HMM-based Synthesis [11, 12] and Speaker Adaptation [13].

In this paper we analyse the performance of the unsupervised cluster generation module whose output will be used to train average voices of the recorded speaking styles. These styled average voices will be used as a background to add expressivity to other neutral voices using well–known speaker adaptive training algorithms [11] commonly used in the flexible framework of HMM-based synthesis [14].

3. C-ORAL-ROM database

The evaluation presented in this paper is carried out using the C–ORAL–ROM [15] database. This corpus is a multi–language and multi–style database covering a wide spectrum of formal and informal speaking styles, in public and private situations.

All the languages included are Romance (French, Italian, Portuguese and Spanish), with styles ranging from formal to informal, extracted either from the media or from private spontaneous natural speaking.

In this paper, the Spanish formal media styles have been analysed: *news broadcasts, sports, meteorological reports, reportage, talk–shows, scientific press and interviews*. These data have been extracted from media broadcasts of different stations, and they present a great deal of variability in the recording environments and a high number of speakers (124). This results in some speakers uttering only a few short sentences, making them almost irrelevant from a statistical parametrical point of view.

Long recordings in this C–ORAL–ROM corpus have been splitted into medium–length sessions . The number of speakers in each session is variable (between 1 and 9 speakers). The maximum length for a specific speaker in one session is 5 minutes. Table 1 summarises average characteristics of the considered sessions for each speaking style.

Table 1: *Features of the speaking style sessions in the C–ORAL–ROM database (ses. stands for session).*

Style	# ses.	SNR	#spk/ses.	time/ses.
interviews	5	25	4	8 min
meteorology	3	26	1	3 min
news	9	29	6	5 min
reportage	15	29	7	5 min
scientific press	4	27	5	9 min
sports	5	33	4	11 min
talk shows	12	29	5	8 min

4. Speaker Diarization System

Unsupervisedly–generated clusters (or pseudo–speakers) will be used for building styled average voices, using the UPM speaker diarization system described in [3].

The system has been adapted to this task in order to use only one feature stream modelling 19 Mel Frequency Cepstrum Coefficients (MFCC). It carries out speaker segmentation and agglomerative clustering of segments, previously filtered by a Voice Activity Detector.

4.1. Speaker Diarization Results

First, we have evaluated the performance by measuring the Diarization Error Rate (DER). Table 2 show that the system achieves a very low DER (lower than 5%) for the styles with speakers that have prepared their discourse, or have some kind of "prompt" (such as in *interviews*, *meteorology* and *news*). More spontaneous speaking styles (such as *reportage*, *sports*, *scientific press* and *talk shows*) obtain a moderate DER (between 10% and 25%).

However, the purpose of this task is to unsupervisedly generate clusters with high precision (or speaker purity) and enough speech to appropriately train a speaking-style average voice or be used as a specific target voice. In this task, miss errors (MISS) are not as relevant as in other tasks (i.e. close captioning) because it will not degrade the models or the output; high percentage of MISS will just force to collect and process more data. Similarly, False Alarms (FA) are also less relevant than the Speaker Error Rate (SER), since the speech–text time alignment module will recover from those FA errors and also because assigning silent frames to a specific cluster (pseudo–speaker) will just provide an improved silence model (assuming a good performance of the automatic time–alignment module).

Table 2: *Speaker diarization results (%) for each speaking style.*

Style	Miss	FA	SpNsp	SER	DER
interviews	0.00	0.30	0.30	6.60	6.93
meteorology	0.00	0.40	0.40	0.70	1.14
news	0.00	0.30	0.30	4.40	4.66
reportages	0.00	1.40	1.40	22.40	23.78
scientific press	0.00	0.30	0.30	15.70	16.01
sports	1.30	0.20	1.50	11.80	13.27
talk shows	2.30	0.30	2.60	13.20	15.78

4.2. Unsupervised Pseudo–Speakers for Speaking Style Average Voices

In this task, the longer and the purer (lower number of actual speakers) a cluster is, the better it would be that cluster for training a natural identifiable styled voice.

This is why the SER commonly used in speaker diarization tasks [2] is not good enough to evaluate the performance of our cluster segmentation system and other metrics like the recall, precision and F-score for each generated cluster are required to evaluate the available data and purity of each generated pseudo–speaker.

Figure 2 shows the recall as a function of the length of the agglomerated clusters. The recall of a cluster has been estimated as the ratio between the length of the speech associated to the speaker that contributed with more speech frames in that cluster and the the length of the speech in that session for that speaker. We obtain a high recall (higher than 70%) considering that it is an unsupervised task over realistic data recorded from the media, not using any phonetic transcription. The recall clearly decreases for clusters smaller than thirty seconds ($t <= 30$), suggesting that thirty is a minimum threshold for this task.

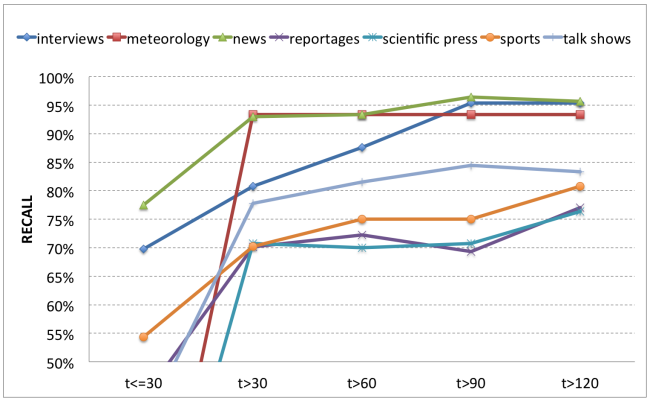

Figure 2: *Recall (%) as a function of the size of the generated clusters (in seconds) for each speaking style*

Figure 3 shows the precision as a function of the size of the clusters. High precision scores are obtained

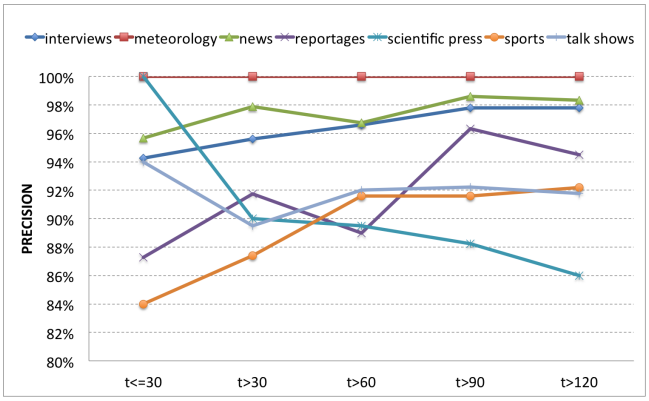

Figure 3: *Precision (in %) as a function of the size of the clusters (in seconds) for each speaking style*

for all the speaking styles (greater than 85% for most lengths). *Meteorology* precision could be considered as trivial since there is only one speaker in each session; however, the diarization system has no prior information about the real number of speakers in each session and nevertheless it has been able to guess that there is only one speaker using purely–unsupervised techniques. *Interviews* and *news* obtain more robust precision scores than the other speaking styles. Background music when certain speakers are talking in *reportage* and *scientific press* sessions explains the lower precision for the longest clusters (especially in case of *scientific press*. This fact confirms the need of a speech/music segmentation module [6] (as proposed in the g framework) which, combined with the VAD module, will filter speech data before diarization.

F–measure results shown in Figure 4 combine both recall and precision in one performance metric. When a cluster is longer than thirty seconds, recall is higher than 75% for every speaking style. This result confirms the thirty–seconds threshold for selecting a cluster. In addition to this, it has been verified that those clusters with high scores correspond to professional speakers (interviewers, journalists, etc) or speakers that are used to speak in the media or in public (such as politicians). On the contrary, *sports* is the most spontaneous style and most of the speakers (excluding the leading journalist) are not used to talk in the media, making them more difficult to diarize.

From these results we have shown that the system performance is significantly higher when the speakers have prepared their discourses than when the spontaneity is higher.

We have carried out an experiment to measure the correlation between the performance of the system and the SNR or the number of speakers in each session: the performance is weakly affected by these two features (Pearson correlation coefficients lower than 0.1%).

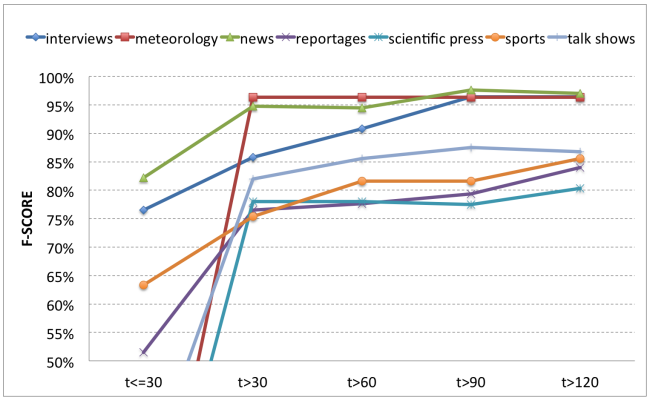

Figure 4: *F-SCORE (in %) as a function of the length of the generated clusters (in seconds) for each speaking style*

It is expected that the precision (or speaker purity) of the unsupervisedly–generated pseudo–speakers would be good enough to build accurate styled average voices with a high similarity with the original speaker and style. In further research it will be necessary to perceptually evaluate the similarity of the synthetic voices with the original speakers, depending of the precision scores of the clusters used in the building process of every speaking style average voice.

5. Conclusion

In this paper we have defined the architecture of the unsupervised multi–speaker voice building framework in *The Simple4All project*. We have shown that the high performance achieved by our speaker diarization system in terms of DER (average of 12%).

As DER is not the best way to evaluate the quality of the pseudo–speakers generated by the system, we have analysed quality in terms of recall, precision and F-score. We have shown that speaking styles with speakers used to talking in public (interviewers, politicians, ect.) favour generating good pseudo–speaker clusters with high precision and recall (higher than 90%). Similarly, speaking styles in which the speakers typically use a certain prompt or prepare their discourse, obtain higher F-scores (90% vs. 80%) than the pseudo-speakers with more spontaneous speaking styles (sports, reportage and talk shows).

Finally, it will be necessary to carry out a perceptual evaluation of the accuracy of the speaking style average voices and the similarity of generated expressive synthetic voices of the target speakers.

6. Acknowledgement

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 MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects. Jaime Lorenzo has been funded by Universidad Politecnica de Madrid under research grant SBUPM-QTKTZHB. Authors also thank the other members of the Speech Technology Group and Simple4All project for the continuous and fruitful discussion on these topics.

7. References

[1] R. Clark and S. King. (2012, March). [Online]. Available: http://simple4all.org

[2] X. Anguera, S. Bozonnet, N. W. D. Evans, C. Fredouille, O. Friedland, and O. Vinyals, "Speaker diarization : A review of recent research," *IEEE Transactions On Audio, Speech, and Language Processing, February 2012, Volume 20, NÂ2, ISSN: 1558-7916*, 05 2011.

[3] J. M. Pardo, R. Barra-Chicote, R. San-Segundo, R. de Cordoba, and B. Martinez-Gonzalez, "Speaker diarization features: The upm contribution to the rt09 evaluation," *IEEE Trransactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 426–435, 2012.

[4] C. Chen, K. Filali, and J. Bilmes, "Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases," in *Proceedings ICSLP 2002*, Sep. 2002, pp. 241–244.

[5] S. Meignier and T. Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, Dallas (Texas, USA), mar 2010.

[6] A. Gallardo and R. San-Segundo, "Upm-uc3m system for music and speech segmentation," in *Jornadas de Tecnologa del Habla FALA 2010*, November 2010.

[7] S. Young, G. Everman, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.1*, Dec. 2001.

[8] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," *Sun Microsystems Technical Report*, no. TR-2004-139, Nov. 2004. [Online]. Available: http://research.sun.com/techrep/2004/abstract-139.html

[9] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proceedings of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Jan. 2011.

[10] K. Prahallad and A. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE Transactions on Speech, Audio & Language Processing*, vol. 19, no. 5, pp. 1444–1449, 7 2011.

[11] J. Yamagishi, "Average-voice-based speech synthesis," Ph.D. dissertation, Tokyo Institute of Technology, 2006.

[12] J. Yamagishi, T. Kobayashi, S. Renals, S. King, H. Zen, T. Toda, and K. Tokuda, "Improved average-voice-based speech synthesis using gender-mixed modeling and a parameter generation algorithm considering GV," in *Proceedings of 6th ISCA Workshop on Speech Synthesis*, Aug. 2007.

[13] 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," *IEEE Transactions on Speech, Audio & Language Processing*, vol. 17, no. 1, pp. 66–83, 1 2009.

[14] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings 6th ISCA Workshop on Speech Synthesis (SSW-6)*, Aug. 2007.

[15] E. Cresti, F. B. do Nascimento, A. M. Sandoval, J. Veronis, P. Martin, and K. Choukri, "The c-oral-rom corpus a multilingual resource of spontaneous speech for romance languages," in *Proceedings of LREC*, 2004.