Towards an Unsupervised Speaking Style Voice Building Framework

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proc. Interspeech 2012

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Towards an Unsupervised Speaking Style Voice Building Framework:
multi–style speaker diarization

J. Lorenzo-Trueba1, B. Martinez-Gonzalez1, V. Lopez–Ludeña1,
R. Barra-Chicote1, J. Ferreiros1, J. Yamagishi2, J.M. Montero3

1Speech Technology Group, ETSIT, Universidad Politecnica de Madrid, Spain
2Centre 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 sponta-
neous and styled speech. This paper proposes an architec-
ture to deal with realistic recordings and carries out some
experiments on unsupervised speaker diarization. In or-
der 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 di-
arization, 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 neu-
tral or expressive voice without needing to record a new
speaker under the target expressivity conditions, main-
taining 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 con-
trol 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 mod-
els.

Another goal of Simple4all is to create the most
portable speech synthesis system possible: one that could
be automatically (or with limited manual supervision) ap-
plied to many domains and tasks, which implies dealing
with a wide variety of expressive situations and do-
 mains. In order to use speech collected from the media
or from media sharing sites, speech synthesis sys-
tems must be robust to the variation of the acoustic and
environmental conditions. The system must 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 diariza-
tion techniques for speaker–turn segmentation will allow
the system creating homogeneous voices from heteroge-
nous recordings, because the number of speakers would
be automatically estimated in a fully unsupervised way,
and language–independent diarization techniques auto-
matically 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 record-
ings. 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, ra-
dio, pod-casts, audio–books, etc.) we define an unsuper-
vised voice–building framework that will allow develop-
ing multiple voices with different speaking styles. The
architecture shown in Figure 1 will minimise the manual
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>
<thead>
<tr>
<th>Style</th>
<th># ses.</th>
<th>SNR</th>
<th>#spk/ses.</th>
<th>time/ses.</th>
</tr>
</thead>
<tbody>
<tr>
<td>interviews</td>
<td>5</td>
<td>25</td>
<td>4</td>
<td>8 min</td>
</tr>
<tr>
<td>meteorology</td>
<td>3</td>
<td>26</td>
<td>1</td>
<td>3 min</td>
</tr>
<tr>
<td>news</td>
<td>9</td>
<td>29</td>
<td>6</td>
<td>5 min</td>
</tr>
<tr>
<td>reportage</td>
<td>15</td>
<td>29</td>
<td>7</td>
<td>5 min</td>
</tr>
<tr>
<td>scientific press</td>
<td>4</td>
<td>27</td>
<td>5</td>
<td>9 min</td>
</tr>
<tr>
<td>sports</td>
<td>5</td>
<td>33</td>
<td>4</td>
<td>11 min</td>
</tr>
<tr>
<td>talk shows</td>
<td>12</td>
<td>29</td>
<td>5</td>
<td>8 min</td>
</tr>
</tbody>
</table>

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].

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.

<table>
<thead>
<tr>
<th>Style</th>
<th>Miss</th>
<th>FA</th>
<th>SpNsp</th>
<th>SER</th>
<th>DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>interviews</td>
<td>0.00</td>
<td>0.30</td>
<td>0.30</td>
<td>6.60</td>
<td>6.93</td>
</tr>
<tr>
<td>meteorology</td>
<td>0.00</td>
<td>0.40</td>
<td>0.40</td>
<td>0.70</td>
<td>1.14</td>
</tr>
<tr>
<td>news</td>
<td>0.00</td>
<td>0.30</td>
<td>0.30</td>
<td>4.40</td>
<td>4.66</td>
</tr>
<tr>
<td>reportages</td>
<td>0.00</td>
<td>1.40</td>
<td>1.40</td>
<td>22.40</td>
<td>23.78</td>
</tr>
<tr>
<td>scientific press</td>
<td>2.30</td>
<td>0.30</td>
<td>2.60</td>
<td>13.20</td>
<td>15.78</td>
</tr>
<tr>
<td>sports</td>
<td>1.30</td>
<td>0.20</td>
<td>1.50</td>
<td>11.80</td>
<td>13.27</td>
</tr>
<tr>
<td>talk shows</td>
<td>2.30</td>
<td>0.20</td>
<td>2.60</td>
<td>11.80</td>
<td>13.27</td>
</tr>
</tbody>
</table>

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 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 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.

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

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.

**Figure 4: F–measure as a function of the size of the generated clusters (in seconds) for each speaking style**

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, etc.) 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-QTKTZH. 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