A Comparison of Open-Source Segmentation Architectures for Dealing with Imperfect Data from the Media in Speech Synthesis

A. Gallardo-Antolín[1], *J. M. Montero*[2], *S. King*[3]

[1]Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain
[2]Speech Technology Group, ETSIT, Universidad Politécnica de Madrid, Spain
[3]The Centre for Speech Technology Research, University of Edinburgh, UK

gallardo@tsc.uc3m.es, juancho@die.upm.es, Simon.King@ed.ac.uk

Abstract

Traditional Text-To-Speech (TTS) systems have been developed using especially-designed non-expressive scripted recordings. In order to develop a new generation of expressive TTS systems in the Simple4All project, real recordings from the media should be used for training new voices with a whole new range of speaking styles. However, for processing this more spontaneous material, the new systems must be able to deal with imperfect data (multi-speaker recordings, background and foreground music and noise), filtering out low-quality audio segments and creating mono-speaker clusters. In this paper we compare several architectures for combining speaker diarization and music and noise detection which improve the precision and overall quality of the segmentation.

Index Terms: diarization, audio segmentation, expressive text-to-speech, media recordings

1. Introduction

The growing interest in the extensive use of TTS systems in different domains and tasks requires building new voices with richer expressivity and speaking styles for many languages in an efficient way. One of the aims of the Simple4all Project [1] is to create universal speech synthesis systems which could be developed automatically (or with limited manual supervision) for specific applications. A key point is the need of the appropriate speech data for the training of new voices with the desired characteristics. Nowadays, a large amount of data available on Internet and the media makes possible to easily collect speech recordings which, however, are not usually fully annotated. These partially annotated or 'found' data must be preprocessed in such a way that they could be suitable for the generation of new voices. Audiobooks are a common choice for this purpose because of their reasonable quality, the presence of a single speaker [2] and rich prosody [3], although other issues such as the segmentation of large audio files into manageable chunks [3] or the speech and text alignment must be addressed [2]. Other resources (broadcast programs, meetings recordings, etc.) produce other challenges, such as multiple speakers and worse recording conditions (background or foreground noise and music) which can negatively affect the quality of the synthesized voice, as pointed in [4].

This paper is focused on the development and analysis of a preprocessing system for unsupervised selection of high quality speech data from media recordings such as radio programmes. For building synthetic voices, the selected audio segments must contain only clean speech (without noise, channel distortions or music) from a single speaker. The main components of the preprocessing system are an audio segmenter for discriminating clean speech from imperfect data (music or noise) and a speaker diarizer for splitting the audio stream into mono-speaker segments.

In order to use as few meta-information and labelling as possible, the speaker diarizer is based on an unsupervised algorithm and audio segmenter models are trained in advance from non-English recordings, although the system is tested on 1-day of BBC Radio 4 programmes.

The main questions we try to answer in this study are:

- Is it worth using an audio segmenter for filtering out imperfect data prior to the speaker diarizer?
- Is it possible to estimate the quality of the speaker diarization clusters from the output of the audio segmenter?

This paper is organized as follows: Section 2 describes two alternative architectures for the preprocessing system. Section 3 presents the database and performance measures used in the experimentation. Section 4 describes the experiments and results to end with some conclusions and ideas for future work in Section 5.

2. System description

The system is composed of two main stages: speech extraction and speech selection. The objective of the first stage is to obtain a set of mono-speaker diarization clusters from a multi-speaker recording with some noise or music segments, sometimes mixed with speech. The aim of the second stage is to choose the more suitable clusters for building synthetic voices, according to the following criteria: quality speech must be as high as possible and the selected clusters must contain enough speech material to train new voices (for example, 500 seconds). Two different architectures combining a speaker diarizer and an audio segmenter will be considered.

2.1. The speaker diarizer and the audio segmenter

We have used the open-source Shout speaker diarizer [5]. It is based on a non-supervised iterative segmentation-clustering algorithm, the Bayesian Information Criterion (BIC) as the cluster-merging criterion and Gaussian Mixture Models (GMM) for modelling the speaker clusters at each iteration [6]. The software allows setting the initial number of clusters (which should be larger than the real number of speakers, which is unknown) and using several feature streams: the primary one is based on Mel-Frequency Cepstrum Coefficients (MFCC) with

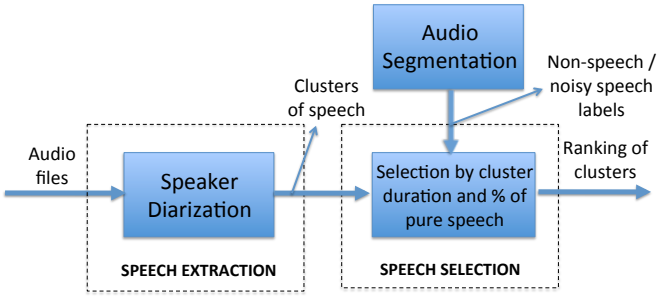

Figure 1: *Block diagram of the SD+AS preprocessing system.*

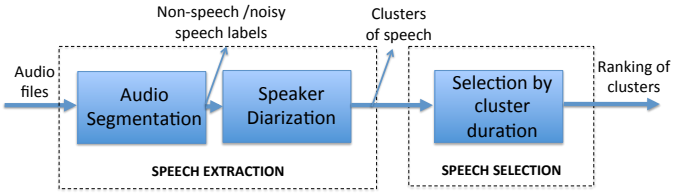

Figure 2: *Block diagram of AS+SD preprocessing system.*

the log-energy and the first derivatives as extracted by Shout itself, and an optional second stream with prosodic information which, in our case, is computed by openSMILE [7].

The audio segmentation module is a non-hierarchical system based on Hidden Markov Models (HMM) [8] and it was initially developed for the segmentation of broadcast news audio documents into five different acoustic classes: clean speech, music, speech with noise (or with several overlapping speakers), speech with music and others. The features (means and standard deviations of 15 MFCCs, log-energy, their corresponding first and second derivatives and 12 CHROMA coefficients [9]) are computed over 1-second frames with an overlap of 0.5 seconds. The system was previously trained on a subset of a Catalan/Spanish broadcast news database which includes around 87 hours of audio. The database consists of broadcast news audio from the 3/24 Catalan TV channel, which was recorded by the TALP Research Center from the Universitat Politècnica de Catalunya, and was manually annotated by Verbio Technologies, in the framework of the Tecnoparla project funded by the Generalitat de Catalunya. Feature extraction is performed by openSMILE [7], and HTK [10] is used for both training and decoding.

2.2. System architectures

We have developed two different architectures for the preprocessing system, as depicted in Figure 1 and Figure 2.

In the first case (denoted as *SD+AS* and represented in Figure 1), the speech extraction process is performed by directly diarizing the raw input audio without any previous processing. Note that, in this approach, the output clusters produced by the diarizer may contain not only pure speech but also noisy speech and even music and other non-speech acoustic events. For this reason, in the speech selection stage, clusters must be chosen taking into account both duration and speech quality. The speech quality of each cluster is estimated as a function of the percentage of pure speech frames detected by the audio segmenter system. In this sense, this parameter can be considered as a kind of quality measure. Obviously, the audio segmentation process is not perfect and some errors may occur. Nevertheless, in the experimentation, we have observed a strong correlation between the percentage of clean speech detected into a cluster and the speaker precision into this cluster. Alternatively, a strong correlation between the cluster quality and the percentage of speech with noise detected has been also observed.

In the second approach (denoted as *AS+SD* and represented in Figure 2), the audio data is previously processed by the audio segmenter, in such a way that segments labelled by this subsystem as "music", "others", "speech with music" or "speech with noise" (in general, non-speech sounds and noisy speech) are filtered out prior to the diarization process. The hypothesis behind this second approach is that the performance of the diarizer will be better if it only processes clean speech data (or at least with a less percentage of contaminated speech or non-speech sounds). As the resulting clusters produced by the speech extraction step are assumed to be composed by only clean speech, in the speech selection stage, clusters are simply chosen according to their duration.

3. Database and performance measures

The database used in this study consists of 35 radio programs of different genres and durations (from 2 minutes to 1 hour long) corresponding to 1-day recordings of BBC Radio 4 (659 speakers). The database was manually transcribed and annotated at speaker level. However, additional information about the presence of music, noise or noisy/low quality speech is not labelled.

As the final objective of the preprocessing system is to obtain mono-speaker speech clusters with as much quality as possible (and taking into account that the labelling concerning to non-speech/noisy speech segments is not available), the performance of the system is measured in terms of precision (at both, speaker and audio file level), although recall and $F_{0.5}$ measure (weighting precision twice as much as recall) are also indicated. Note that the information about the number and identity of speakers are only used for computing the corresponding performance measures.

4. Experiments

4.1. Preliminary experiments on the speech extraction stage of the SD+AS system

First of all, a set of preliminary experiments was performed on the speech extraction stage of the *SD+AS* system in order to adjust some free parameters of the diarizer: the maximum number of initial clusters ("MC") and the use of prosodic information (pitch and intensity) as a second feature stream.

Table 1 contains the results of these experiments in terms of precision, recall and $F_{0.5}$ averaged over all the files of the database. In the experiments with a variable MC (rows 3 to 5), a different number of initial clusters was set as a function of the audio file length, as a reasonable assumption is that duration is highly correlated with the number of speakers contained into the given audio recording. In particular, the initial number of clusters was set according to Table 2.

A variable MC improves significantly the precision of the system whereas the recall slightly decreases, especially for long files. Regarding prosodic information, it degrades the performance of the system in terms of average precision. For this reason, for the rest of the experiments with the *SD+AS* system, we have only used the first feature stream and a variable MC.

Table 1: *SD+AS system: performance of the extraction stage.*

MC	Streams (weights)	Precision	Recall	$F_{0.5}$
16	1 str.: MFCC	70.97%	85.34%	0.73
Variable	1 str.: MFCC	75.35%	84.37%	0.77
Variable	2 str.: MFCC (0.8) pitch+intensity (0.2)	71.47%	86.13%	0.74
Variable	2 str.: MFCC (0.9) pitch+intensity (0.1)	72.19%	86.54%	0.74

Table 2: *Number of initial clusters as a function of the audio duration.*

Duration of the audio file	Number of initial clusters (MC)
dur <15 min	16
15 min <dur <45 min	24
dur >45 min	32

4.2. Preliminary experiments on the speech extraction stage of the AS+SD system

In this case, several experiments were performed by varying the insertion penalty ("IS") in the audio segmenter and the maximum number of initial clusters ("MC") and the use of two feature streams in the diarizer. Results of these experiments are shown in Table 3.

The insertion penalty controls the number of insertions in the resulting segmentation, so smaller values of this parameter tend to produce longer segments. As it can be observed in Table 3, both, precision and recall, do not suffer large variations with IS values in the range from -10 to -20. With respect to the number of initial clusters and the use of two feature streams, the behaviour of the system is the same as for the *SD+AS* case, so for the rest of experiments the configuration will be: IS = -10, MC = Variable and 1 stream (MFCC).

Table 3: *AS+SD system: performance of the extraction stage.*

IS	MC	Streams (weights)	Precision	Recall	$F_{0.5}$
-20	16	1 str.: MFCC	77.47%	76.84%	0.76
-15	16	1 str.: MFCC	77.59%	75.33%	0.76
-10	16	1 str.: MFCC	78.29%	75.50%	0.77
-10	Var.	1 str.: MFCC	82.56%	77.33%	0.80
-10	Var.	2 str.: MFCC (0.8) pitch+int.. (0.2)	76.81%	75.16%	0.76
-10	Var.	2 str.: MFCC (0.9) pitch+int. (0.1)	77.73%	75.77%	0.77

4.3. Comparison between the speech extraction stage of SD+AS and AS+SD systems

Table 4 contains the precision, recall and $F_{0.5}$ measures for the best configuration of the *SD+AS* and *AS+SD* systems averaged over the radio programs of the same genre and over all programs. For comparison purposes, it is also shown the simplest case in which it is assumed that each complete audio file contains only one speaker (column label as "Without preprocessing").

From the results, it is clear that the system without preprocessing does not produce high precision speech clusters, so it is necessary to use either of the two preprocessing systems considered. The *AS+SD* system outperforms the *SD+AS* one for all the genres in terms of precision and $F_{0.5}$, although an increase in precision entails a decrease in recall. Any case, it seems that the speech extraction stage of *AS+SD* produces better quality clusters than *SD+AS*.

Also, it is worth mentioning that there are important variations in precision, recall and $F_{0.5}$ with respect to the genre of the radio program. As expected, on average, it is more difficult to extract useful speech clusters from programs with a large number of speakers (i.e. with longer duration in most of the cases) or containing music, other kind of non-speech sounds or noisy speech (as for example, drama-sitcoms) than from programs with few speakers and recorded in studio conditions (as for example, weather bulletins or drama-readings). However, this fact does not imply that the first kind of programs should be directly discarded, because they could contain several high quality speakers. For this reason, the speech selection stage must be performed at cluster level and not at file level.

4.4. Experiments with the complete SD+AS system

Table 5 shows the ranking of the speech clusters longer than 500 seconds provided by the *SD+AS* system. The last column contains the new ranking position of the given cluster after taking into account its clean speech content (in this experiment, clusters with a percentage of pure speech less than 90% are discarded). Note that in this case, the output of the audio segmentation module is used as a kind of quality measure.

As it can be observed, when clusters are selected only according to their length, 6 out of 17 clusters have a precision lower than 85% and therefore, presumably, their quality is low. The audio segmenter is capable of discarding some of these files (dark gray rows), but 4 of them are still on the first positions of the list (light gray rows), because the diarizer performance is affected by the presence of music and other noises, in such way that in some occasions (see, for example, the cluster "SPK2") it creates small clusters containing music and other non-speech sounds whereas it merges the speech of various speakers in a large only-speech cluster. In these cases, the audio segmenter labels the entire cluster as clean speech, and so, it is not capable of detecting it as a low quality cluster.

4.5. Experiments with the complete AS+SD system

The list of clusters with a duration above 500 seconds obtained with the *AS+SD* system is shown in Table 6. In this case, the clusters were ranked only by decreasing length, as it is assumed that non-speech sounds and noisy speech segments have been already detected and eliminated by the audio segmentation system in the speech extraction stage. For comparison purposes, the two last columns indicate, respectively, the position of the same speaker in the ranking of the *SD+AS* system (last column of Table 5) and its relative precision improvement with respect to the *SD+AS* system.

As it can be observed, all the clusters in the top of the list generated by the *AS+SD* system have a precision greater than 88%. In addition, the average relative precision improvement with respect to the *SD+AS* system is around 2.5%. These results suggest that clusters selected by *AS+SD* have higher quality than the ones provided by *SD+AS* and therefore, are most suitable for the generation of new synthetic voices from them.

Table 4: *Experimental results for the SD+AS and AS+SD systems. Performance measures for the case in which no preprocessing is used are also included for comparison purposes.*

Genre	Without preprocessing			SD+AS; MC = variable; 1 str. (MFCC)			AS+SD; IS = -10; MC = variable; 1 str. (MFCC)		
	Prec.	Rec.	$F_{0.5}$	Prec.	Rec.	$F_{0.5}$	Prec.	Rec.	$F_{0.5}$
Weathers - Bulletins	76.11%	100%	0.79	94.48%	97.17%	0.95	96.62%	92.64%	0.96
News	28.14%	100%	0.32	69.61%	94.64%	0.73	77.81%	84.94%	0.79
Drama - Readings	57.19%	100%	0.61	89.04%	98.76%	0.91	91.52%	87.04%	0.91
Drama - Sitcoms/Soaps	26.49%	100%	0.31	48.60%	53.16%	0.49	63.05%	39.54%	0.56
Factual - Magazine & Reviews	40.29%	100%	0.45	75.55%	81.32%	0.76	82.33%	72.05%	0.79
Factual - Discussion & Talk	24.10%	100%	0.28	90.48%	92.30%	0.91	90.95%	89.89%	0.91
Factual - Documentary	40.86%	100%	0.46	78.18%	86.13%	0.79	86.43%	74.96%	0.83
Entertainment - Games	39.83%	100%	0.45	87.36%	86.33%	0.87	93.76%	66.75%	0.86
Total	40.08%	100%	0.44	75.35%	84.37%	0.77	82.56%	74.33%	0.80

Table 5: *Ranking of the speech clusters selected by the SD+AS system.*

Speaker name	Cluster duration (s)	Precis.	Recall	% clean speech	Ranking after AS
SPK1	959.86	88.93 %	92.64 %	100.0%	1
SPK2	909.13	28.39 %	99.96%	95.59%	2
SPK3	734.00	92.29%	97.65%	100.0%	3
SPK4	671.53	80.32%	99.47%	93.72%	4
SPK5	654.92	97.11%	98.06%	98.50%	5
SPK6	641.96	95.29%	95.58%	98.14%	6
SPK7	633.92	21.25%	90.29%	14.65%	Discarded
SPK8	607.05	44.91%	66.04%	70.27%	Discarded
SPK9	596.20	94.45%	98.67%	97.55%	7
SPK10	590.85	25.03%	100.0%	91.02%	8
SPK11	577.81	89.51%	87.82%	97.09%	9
SPK12	572.56	81.29%	72.82%	100.0%	10
SPK13	571.13	98.26%	100.0%	99.54%	11
SPK14	549.30	87.62%	93.98%	100.0%	12
SPK15	542.72	96.97%	86.31%	100.0%	13
SPK16	533.00	94.15%	99.98%	85.38%	Discarded
SPK17	518.25	96.82%	100.0%	99.56%	14

Table 6: *Ranking of the speech clusters selected by the AS+SD system.*

Speaker name	Cluster duration (s)	Precis	Recall	Ranking SD+AS (after AS)	Relat. Precis. Improv.
SPK1	798.78	98.85%	85.69%	1	11.15%
SPK3	722.45	93.01%	96.87%	3	0.78%
SPK5	655.27	97.83%	98.83%	5	0.74%
SPK15	621.86	97.00%	98.93%	13	0.03%
SPK6	596.98	97.49%	90.93%	6	2.31%
SPK13	568.52	98.26%	99.54%	11	0.00%
SPK4	554.07	88.96%	90.90%	4	10.76%
SPK14	545.88	88.27%	94.09%	12	0.74%
SPK9	532.34	92.41%	86.20%	7	-2.15%
SPK17	501.95	97.75%	97.79%	14	0.96%

5. Conclusions

In this work, we have developed a system for the unsupervised extraction of high-quality speech from complex audio recordings, which can be used as a preprocessing stage for selecting speech material useful for building new voices with new speaking styles and expressivity. Two different architectures for this system combining a speaker diarizer and and audio segmenter have been evaluated.

The performance of the speaker diarization system improves when it is applied over clean (or at least partially clean) speech, so filtering out non-speech/noisy speech segments prior to the application of the diarizer is useful for this task, even when the audio segmentation system was not trained on the testing audio data. In summary, it seems that the best architecture for the preprocessing system is the one involving an audio segmentation step followed by a diarization stage.

The output of the audio segmenter can be used as a quality measure of the resulting speech clusters when they contain non-speech acoustic events or speech contaminated with noise or music as the percentage of clean speech detected in a given cluster has a strong correlation to its precision. However, the audio segmenter is not able to detect the case in which the cluster is composed of speech from several speakers.

The future lines of research includes training/adapting new voices with new speaking styles and expressivity from the speakers selected by the preprocessing system described in this work (some samples of prosodic HTS synthesis are already available at [1]). Also, we plan to improve the audio segmentation module by using features derived from the decomposition of the audio spectrum into speech and non-speech components provided by the application of Non-Negative Matrix Factorization (NMF) in a semi-supervised way [11] and [12].

6. Acknowledgements

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 Nº 287678. It has also been supported by EPSRC Programme Grant grant, no. EP/I031022/1 (Natural Speech Technology, NST) and Spanish Government grants TEC2011-26807 and DPI2010-21247-C02-02. The data used in this research was kindly provided by BBC R&D as part of NST. The authors thank TALP-UPC for providing data for training the audio segmenter and Rubén San Segundo and Beatriz Martínez for providing some pieces of software for these experiments.

7. References

[1] R. Clark and S. King. [Online]. Available: http://simple4all.org, accessed on Mar 2014.

[2] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, "TUNDRA: A multilingual corpus of found data for tts research created with light supervision," in *Proc. Interspeech 2013*, Lyon, France, 2013.

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

[4] R. Karhila, U. Remes, and M. Kurimo, "Hmm-based speech synthesis adaptation using noisy data: Analysis and evaluation methods," in *Proc. of ICASSP 2013*, Vancouver, Canada, 2013, pp. 6930–6934.

[5] M. Huijbregts, "Large vocabulary continuous speech recognition toolkit (Shout)." [Online]. Available: http://shout-toolkit.sourceforge.net/index.html, accessed on Mar 2014.

[6] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. of ASRU 2003*, Virgin Islands, USA, 2003.

[7] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM), ACM*, Florence, Italy, 2010, pp. 1459–1462.

[8] A. Gallardo-Antolín and R. San-Segundo, "UPM-UC3M system for music and speech segmentation," in *Proc. VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop (FALA 2010)*, Vigo, Spain, 2010, pp. 421–424.

[9] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: using chroma-based representations for audio thumbnailing," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001, pp. 15–18.

[10] S. Young *et al.*, *HTK-Hidden Markov Model Toolkit (Ver 3.4)*. Cambridge, MA: Cambridge Univ., 2009.

[11] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization-based compensation of music for automatic speech recognition," in *Proc. of Interspeech 2010*, Makuhari, Japan, 2010, pp. 717–720.

[12] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, and G. Rigoll, "Non-negative matrix factorization as noise-robust feature extractor or speech recognition," in *Proc. of ICASSP 2010*, Dallas, USA, 2010, pp. 4562–4565.