Benefits of prior acoustic segmentation for automatic speaker segmentation

Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-François Bonastre, Laurent Besacier

BENEFITS OF PRIOR ACOUSTIC SEGMENTATION FOR AUTOMATIC SPEAKER SEGMENTATION

Sylvain Meignier [(1)], Daniel Moraru [(2)], Corinne Fredouille [(1)], Laurent Besacier [(2)], Jean-François Bonastre [(1)]

[1] LIA-Avignon - BP1228 - 84911 Avignon Cedex 9 – France

[2] CLIPS-IMAG (UJF & CNRS) - BP 53 - 38041 Grenoble Cedex 9 - France

[1] (sylvain.meignier,corinne.fredouille,jean-francois.bonastre)@lia.univ-avignon.fr,

[2] (daniel.moraru,laurent.besacier)@imag.fr

ABSTRACT

This paper investigates the interest of a segmentation in acoustic macro classes (like gender or bandwidth) as a front-end processing for the segmentation/diarization task. The impact of this prior acoustic segmentation is evaluated in terms of speaker diarization performance in the particular context of NIST RT'03 evaluation (done on HUB4 broadcast news corpora). Rarely discussed in the literature, this work shows that the application of a prior acoustic segmentation, in a similar way to the automatic speech recognition task, may be very useful to the speaker segmentation task. The experiments were conducted using two different kinds of speaker segmentation systems developed individually by the LIA and CLIPS laboratories in the framework of the ELISA consortium. For both systems, improvement was observed when combined with the prior acoustic segmentation. However, a larger impact, in terms of performance, is observed on the ascending/HMM approach based LIA system compared to the speaker turn detection based CLIPS system.

1. INTRODUCTION

Speaker diarization (or segmentation) is a new speech processing task resulting from the increase in the number of multimedia documents that need to be properly archived and accessed. One key of indexing can be speaker identity. The goal of speaker diarization is to segment a N-speaker audio document in homogeneous parts containing the voice of only one speaker (also called speaker change detection process) and to associate the resulting segments by matching those belonging to the same speaker (clustering process). Generally, no *a priori* information is available on the number of speakers involved in the conversation as well as on the identity of the speakers.

This paper is focused on the acoustic segmentation task, which was mainly introduced to help automatic speech recognition (ASR) systems within the special context of broadcast news transcription. Indeed, at the beginning, one of the main objective of the acoustic segmentation was to provide ASR systems with an acoustic event classification allowing to discard nonspeech signal (silence, music, …), to adapt ASR acoustic models to some particular acoustic environments (like speech over music, telephone speech) or simply to speaker gender [1][2][1]. Many papers were dedicated to this particular issue and to the evaluation of acoustic segmentation in the context of ASR task. Nevertheless, rarely discussed in the literature, acoustic segmentation may be useful for other tasks linked to broadcast news corpora. In this sense, the aim of this paper is to investigate the impact of acoustic segmentation when applied as a prior segmentation for the particular task of speaker segmentation. The LIA macro class acoustic segmentation process is combined with two different speaker segmentation systems, developed individually by the CLIPS and the LIA laboratories (in the framework of the ELISA consortium), which exhibit two different segmentation strategies.

The experiments were conducted in the framework of the NIST 2003 Rich Transcription (RT) evaluation campaign[2]. This project is sponsored in part of the DARPA Effective Affordable Reusable Speech To Text (EARS) Program. The EARS research effort is dedicated to developing powerful speech transcription technology that provides rich and accurate transcripts. It includes speech transcription but also acoustic segmentation, speaker indexing, disfluency detection induced by spontaneous speech (hesitations, self repairs, word fragments…), etc. This will help machines to perform much better on detecting, extracting, summarizing, and translating important information. EARS is focusing on natural, unconstrained human-human speech from broadcast and foreign conversational speech in multiple languages.

Section 2 presents the LIA macro class acoustic segmentation system, based on a hierarchical strategy while *section 3* provides a brief description of both speaker segmentation systems developed at the LIA and the CLIPS labs. The experimental context and the speaker segmentation system performance are presented and discussed in *section 4*. Finally, some conclusions and perspectives are given in *section 5*.

2. ACOUSTIC MACRO CLASS SEGMENTATION

Acoustic macro class segmentation is necessary to discard nonspeech signal (like music, silence, …) or to adapt acoustic models to specific acoustic environments (telephone speech, speech over music, …). This is especially true when handling broadcast news data with the aim of automatically transcribing speech for instance. Basic speech/nonspeech detection is also useful for speaker segmentation task in order to avoid music portions to be automatically labeled as a new speaker. However, acoustic segmentation system may be designed to provide finer

[1] Speaker gender may be considered as a particular acoustic class.

[2] More information could be found at http://www.nist.gov/speech/tests/rt/rt2003/index.htm

classification. For example, gender classification could help the segmentation process, by selecting the appropriate a priori knowledge. In this paper, the prior acoustic segmentation is done at three different levels: speech / nonspeech, clean speech/speech over music/telephone speech[3] and male/female speech.

2.1 Front end processing

The signal is characterized by 39 acoustic features computed every 10 ms on 25 ms Hamming-windowed frames: 12 MFCC augmented by the normalized log-energy, followed by the delta and delta-delta coefficients.

2.2. Hierarchical approach

The system relies on a hierarchical segmentation performed in three successive steps as illustrated in figure 1:

- during the first step, a speech/nonspeech segmentation of signal (representing a show) is performed using "MixS" and "NS" models. The first model represents all the speech conditions while the second one represents the nonspeech conditions. Basically, the segmentation process relies on a frame-by-frame best model search. A set of morphological rules are then applied to aggregate frames and label segments.
- during the second step, a segmentation based on 3 classes — clean speech ("S" model), speech over music ("SM" model) and telephone speech ("T" model) — is performed only on the speech segments detected by the previous segmentation step. All the models involved during this step are gender-independent. The segmentation process is a Viterbi decoding applied on an ergodic HMM, composed, here, of three states ("S", "T", and "SM" models). The transition probabilities of this ergodic HMM are learnt on 1996 HUB 4 broadcast news corpus.
- the last step is devoted to gender detection. According to the labels given during the previous step, each segment will be identified as female or male speech by the use of models dependent on both gender and acoustic classes. "GT-Fe" and "GT-Ma" models represent female and male telephone speech respectively, "GS-Fe" and "GS-Ma" represent female and male clean speech, while "GSM-Fe" and "GSM-Ma" represent female and male speech over music. Two other models, "GDS-Fe" and "GDS-Ma", representing female and male speech recorded over degraded conditions are also used, at this stage, to refine the final segmentation. The segmentation process, described in the previous step, is applied in the same way here.

All the state models mentioned above are diagonal GMMs ([1][3]). Except "NS" and "MixS" models, which are characterized by 1 and 512 Gaussian components respectively, all the other models are characterized by 1024 Gaussian components. They were trained on the 1996 HUB 4 broadcast news corpus.

3. SPEAKER SEGMENTATION SYSTEMS

Two speaker segmentation systems are presented briefly in this section, relying on different segmentation strategies. These speaker segmentation systems were developed individually by the CLIPS and LIA laboratories in the framework of the ELISA consortium [5].

Both systems rely on the LIA speaker recognition system, named AMIRAL [4]. The two systems are based on Gaussian Mixture Models (GMM). More technical information about both speaker segmentation systems may be found in [6][7].

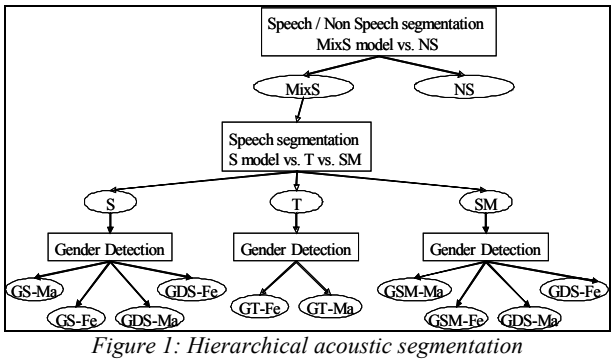

Figure 1: Hierarchical acoustic segmentation

3.1. Speaker-turn-point detection based system

The CLIPS speaker segmentation system is based on a standard approach relying on a speaker turn detection followed by a hierarchical clustering.

The potential speaker turn points are determined using a Bayesian Information Criterion (BIC) approach. A BIC curve is extracted by computing a distance between two 1.75s adjacent windows that go along the signal. Mono-Gaussian models with diagonal covariance matrices are used to model the two windows. A threshold is then applied on the BIC curve to find the most likely speaker change points, which correspond to the local maximum of the curve.

For clustering, diagonal 32 component GMM are used to model the segments. Next, BIC distances are computed between segment models and the closest segments are merged at each step of the algorithm until N segments are left (corresponding to N speakers in the conversation). The number of speakers (N) is estimated using a penalized BIC (Bayesian Information Criterion).

3.2. Ascending/HMM based system

The speaker segmentation system developed by the LIA laboratory relies on an ascending Hidden Markov Modeling (HMM) of the conversation/show. Each state of the HMM characterizes a speaker and the transitions model the changes between speakers.

The HMM model is built iteratively by adding the speakers one by one. In this framework, speaker change detection and clustering are performed concurrently in a single iterative process. At each step of this process a speaker is added, the HMM transition parameters are moved to reflect the new HMM structure and an iterative adaptation process is done. During this adaptation phase the models are adapted (acoustic model adaptation) corresponding to the current segmentation and a new segmentation is computed using Viterbi decoding (speaker change points and speaker clustering refinement). The last phases are repeated until no gain is observed. The final segmentation is achieved when a stop criterion — based on both likelihood of the segmentation and on some heuristics — is reached.

[3] This segmentation may also be referred to as a narrow/wide band speech classification if "speech over music" label is not used and simply considered as "speech" label.

Finally, the resulting speaker segmentation is refined during a last re-segmentation step similar to the previous one except that a different acoustic model (GMM) adaptation algorithm is used because in this case the audio utterances associated with each speaker are longer.

3.3. Combination between prior acoustic segmentation and both speaker segmentation systems

The information yielded by the prior acoustic segmentation is taken into account the same way by the two speaker segmentation systems. Both systems use the prior acoustic segmentation in order to suppress the nonspeech segments. This speech/nonspeech classification is never questioned again during the speaker segmentation process. However, the intrinsic behavior of speaker segmentation processes, especially in the application of ascending/HMM approach, may lead to some deviations of speech/nonspeech frontiers.

Both CLIPS and LIA apply their speaker segmentation system separately on the different classes of acoustic events detected by the acoustic segmentation system. Then, different acoustic class dependent segmentations are merged together to yield a final segmentation.

In both cases, this final segmentation could be consolidated through a last re-segmentation phase, identical to the final refining step of the LIA system, described in 3.2.

4. EVALUATION OF ACOUSTIC SEGMENTATION IMPACT ON SPEAKER SEGMENTATION

This section presents the evaluation protocol used to measure the impact of the acoustic macro class segmentation when combined with speaker segmentation and discusses the experimental results obtained in this framework. Different levels of acoustic segmentation granularity are evaluated on both speaker segmentation systems:

- Speech/nonspeech classification only (S/NS);
- segmentation based on speech/nonspeech and gender detection (S/NS—Gender);
- segmentation based on speech/non speech, gender and telephone/non telephone speech detection (S/NS—Gender—T/NT);
- segmentation based on speech/nonspeech, gender and telephone/clean speech/speech over music/degraded speech (S/NS—Gender—T/S/MS/DS).

For comparison, some speaker segmentation results will be also presented based on an acoustic segmentation, obtained manually and based on speech/nonspeech, gender and telephone/non telephone speech detection (Hand S/NS—Gender—T/NT).

4.1. Evaluation protocol

The experiments described in this paper were conducted in the framework of NIST/RT'03 evaluation campaign[2]. In this context, two separate corpora were available :

- the *Dev* corpus composed of 6 broadcast news shows of 10mn each, recorded in 1998, available to tune the speaker segmentation systems;
- the *Eva* corpus composed of 3 broadcast news shows of 30mn each, recorded in 2001, available for evaluation only.

For both corpora, we manually discarded advertisement portions before any treatment, that explains that results presented here for both speaker segmentation systems do not correspond exactly to official RT'03 results

Speaker segmentation system performance was measured in terms of diarization error rate, according to NIST/RT'03 scoring[2], which takes into account the speaker segmentation error, as well as missed and false alarm speaker errors, directly linked to speech/nonspeech classification errors.

4.2. Performance of the prior acoustic segmentation

Table 1 gives the performance of the prior acoustic segmentation process. The speech/nonspeech segmentation error is around 4.5% (in terms of duration) compared to 4.4% for the best system during NIST RT'03 evaluation campaign. The gender detection error goes from 1.5% for the *Dev* set to 5.5% for the *Eva* set. Thanks to the manual segmentation Hand S/NS—Gender—T/NT, the accuracy of acoustic segmentation system at the level of telephone and non telephone classification is evaluated: less than 0.1% for *Dev* corpus and 3% for *Eva*.

Corpus	Missed Speech Error	False Alarm Speech Error	Gender Error	Telephone / Non telephone Speech error
Dev	2.3%	2.2%	1.5%	0.09 %
Eva	1.8%	2.7%	5.5%	3 %

Table 1: Acoustic segmentation errors on Dev and Eva sets.

4.3. Overall segmentation system results

Table 2 and table 3 present the experimental results obtained when combining speaker segmentation systems (speaker turn detection based and ascending/HMM based approaches) with the acoustic segmentation before and after the re-segmentation step respectively. These results, expressed in terms of diarization error rates, show :

- a large variation in terms of performance between *Dev* and *Eva* corpora, especially for the ascending/HMM approach;
- a gain in performance (the best one) for both speaker segmentation systems when they are combined with the manual acoustic segmentation;
- a significant improvement of ascending/HMM approach results on *Eva* corpus due to the speech/nonspeech, gender and telephone/non telephone segmentations (S/NS—Gender—T/NT) without (from 26.9% to 18.1%) and with (from 26.6% to 14.1%) the re-segmentation phase. On *Dev* corpus, this improvement is visible only after the re-segmentation phase (from 15.5% to 12.8%);
- no real effect on the speaker-turn-based approach regarding the different levels of acoustic segmentation granularity, except on *Eva* corpus after the re-segmentation step for which result improvement can be observed (from 15.7 to 13.9%);
- no improvement or even performance loss when the finest acoustic segmentation (S/NS—Gender—T/S/MS/DS) was involved;
- the interest of the re-segmentation step for both speaker segmentation strategies since best performance may be observed in most of the cases.

4.4. Discussion

Comparing all the different levels of segmentation granularity, the S/NS—Gender—T/NT segmentation seems the most helpful for the speaker segmentation task, especially for the ascending/HMM approach.

	Ascend./HMM approach		Speaker turn approach	
Acoustic segmentation	***Dev***	***Eva***	***Dev***	***Eva***
Hand S/NS-Gender-T/NT	*13.1%*	*15.4%*	*14.1%*	*10.6%*
S/NS	15.4%	26.9%	19.7%	17.1%
S/NS-Gender	14.8%	27.4%	18.7%	18.6%
S/NS-Gender-T/NT	15.1%	18.1%	19.0%	18.2%
S/NS-Gender-T/S/MS/DS	25.9%	30.5%	19.1%	27.6%

Table 2: Diarization error rates for different levels of acoustic granularity on Dev and Eva sets before re-segmentation step.

	Ascend./HMM approach		Speaker turn approach	
Acoustic segmentation	***Dev***	***Eva***	***Dev***	***Eva***
Hand S/NS-Gender-T/NT	*10.8%*	*13.2%*	*15.7%*	*9.4%*
S/NS	15.5%	26.6%	19.3%	15.7%
S/NS-Gender	13%	24.9%	18.2%	15.3%
S/NS-Gender-T/NT	12.8%	14.1%	19.0%	13.9%
S/NS-Gender-T/S/MS/DS	15.6%	14.3%	18.7%	15.1%

Table 3: Diarization error rates for different levels of acoustic granularity on Dev and Eva sets after re-segmentation step.

This point is particularly visible for *Eva* corpus for which 20% of speech time (shared among 2 shows over the 3 available in the corpus) is telephone speech against 7.7% only for *Dev* corpus (mainly present in 1 show over the 6 available).

The difference of behaviors between the two speaker segmentation systems (no real impact on the speaker turn approach) may be directly linked to the strategies involved in each of them. It seems reasonable that the speaker turn approach intrinsically behaves as an acoustic class segmentation system, detecting speaker turns as well as acoustic event changes before the clustering phase. In this sense, the a priori acoustic macro class segmentation becomes useless to improve performance, except for speech/nonspeech detection since speaker turn approach cannot discard nonspeech segments automatically without additional treatment.

As not expected, the finest segmentation, S/NS—Gender—T/S/MS/DS, does not lead to performance gain and may inversely degrade it in some cases. This can be explained by the fact that some speakers may be present under different acoustic classes (speech over music followed by speech only, classical for news presenters or in both clean and degraded speech classes depending on the location of interviews for instance). Since speaker segmentation systems are applied independently on each acoustic class, a same speaker may be split under different labels, leading to an increase in speaker error rates. In the same way, increasing the number of acoustic classes induces much more smaller segments, which may disturb speaker segmentation systems. These effects are however partially overcome thanks to the re-segmentation phase.

Finally, combining both speaker segmentation systems with the manual acoustic segmentation outperforms all the automatic ones. However, it is worth noting again that diarization error rate takes into account both speaker and speech/nonspeech error rates. Regarding speaker error rate only the best speaker segmentation system (after re-segmentation) based on an automatic acoustic segmentation on *Eva* corpus gets 8% against 7.2% for the manual segmentation. This tends to show that segments misclassified by the automatic acoustic segmentation system may be corrected by the re-segmentation step and therefore do not disturb the speaker segmentation process.

5. CONCLUSION

This paper investigated the impact of *prior* acoustic macro class segmentation when it is combined with speaker segmentation. This impact was evaluated in terms of speaker segmentation performance and, more precisely, in terms of speaker diarization error rate, according to NIST/RT'03 scoring. This investigation was conducted on two speaker segmentation systems, exhibiting two different segmentation strategies: a speaker turn detection based system and an ascending/HMM based one, developed individually by the CLIPS and LIA labs respectively. The *prior* acoustic macro class segmentation, presented in this paper, was developed by the LIA lab. It relies on a GMM model based hierarchical segmentation, designed to provide different levels of segmentation granularity (from simple speech/nonspeech detection to gender dependent acoustic classes such as speech over music, degraded speech or telephone speech).

Experiments based on the combination of each speaker segmentation system with the acoustic segmentation were conducted according to the different levels of acoustic segmentation granularity. The results presented in this paper show the benefit of acoustic segmentation for speaker segmentation performance, especially for the ascending/HMM approach, while the speaker turn detection based system seems to be less dependent of an acoustic segmentation.

Further work should study the way of taking benefit of finer acoustic classes such as speech over music or degraded speech for the speaker segmentation task.

6. REFERENCES

[1] P.C. Woodland, "The development of the HTK Broadcast News transcription system: An overview", *Speech Communication*, Vol. 37, pp. 291-299, 2002.

[2] J.L. Gauvain, L. Lamel, and G. Adda. "The LIMSI Broadcast News Transcription System". *Speech Communication*, 37(1-2):89-108, 2002.

[3] T. Hain, and P.C. Woodland, "Segmentation and Classification of Broadcast News audio", *ICSLP'98*, Sydney, Australia.

[4] C. Fredouille, J.-F. Bonastre, and T. Merlin, "AMIRAL: a block-segmental multi-recognizer architecture for automatic speaker recognition," *Digital Signal Processing*, Vol. 10, No. 1-3, January/April/July 2000.

[5] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet for the ELISA consortium, "Overview of the 2000-2001 ELISA consortium research activities," in *2001: A Speaker Odyssey*, pp.67–72, Chania, Crete, June 2001.

[6] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau, "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation". *ICASSP'03*, Hong Kong.

[7] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, "The ELISA consortium approaches in Broadcast News speaker segmentation during the NIST 2003 Rich Transcription evaluation". *Paper submitted at ICASSP'04,* Montreal, Canada.