Automatic Turn Segmentation in Spoken Conversations

Alexei V. Ivanov, Giuseppe Riccardi

Department of Information Engineering and Computer Science,
University of Trento, Povo (Trento), Italy
ivanov@disi.unitn.it, riccardi@disi.unitn.it

Abstract

In this paper we have studied the problem of detecting the spoken turn boundaries in human-human spoken conversations. The automation of this task is essential to enable the analysis, recognition and understanding of the speech transcriptions and dialog structures (e.g. turn taking, dialog act segmentation etc.). The problem formulation is different from previous work on metadata extraction in that we work on the time domain for the detection of boundaries. This approach has the advantage of giving fine grain measures of speech events and does not rely on the automatic speech transcriptions. We have explored applicability of different algorithms for this task and have found that a hidden Markov model combining results of the modulation spectrum analysis and Kullback-Leibler divergence of adjacent signal portions produces the best results. The performance of the algorithms has been evaluated on the Switchboard conversational speech corpus.

Index Terms: spoken turn boundary , spoken dialogs, modulation spectrum, Bayesian information criterion, Kullback-Leibler divergence

1. Introduction

Understanding human conversations is a complex process involving speaker turn segmentation, turn transcription, attribution of a dialog act, detection of key concepts, etc. Automated completion of these tasks is crucial to build machines for supporting humans in complex analysis of large amount of dialog data. In this paper we focus on the automatic detection of turn boundaries in the time domain. Such task is relevant for downstream processes such as automatic speech recognition, turn taking modeling, spoken utterance human annotation, etc. The problem formulation is different from previous work on metadata extraction [1] in that we work on the time domain for the detection of boundaries. This approach has the advantage of giving fine grain measures of speech events and does not rely on the speech transcriptions generated by the Automatic Speech Recognizer (ASR). Our approach is also different from the traditional voice-activity detection (VAD). The classical VAD problem [2] was formulated in the context of speech coding [3, 4] with the purpose to release communication bandwidth when there is no *useful* signal to be transmitted. The definition of the "useful signal" in the classical VAD task is very broad: any signal with the spectral power distribution being far from the uniform is deemed to be worth of reporting as a useful signal.

The typical time duration of spoken turns is in the range of 0.5 – 30 sec. It is acceptable and even beneficial to include leading and trailing silence frames so that downstream ASR can have proper initialization and shutdown. Moreover, it is more important to keep the rate of falsely taken decisions low rather than attempt to achieve the best ratio of correctly classified analysis frames, because each wrongly taken decision would lead to an ASR almost certainly making erroneous recognition of the underlying spoken turn. Since the acquisition is typically done through a close-talking microphone, speech in telephone conversations rarely gets contaminated with a moderate stationary noise, however the non-stationary signal portions are to be expected. Ideally the speech turn segmenter should not pass to a recognizer anything but speech intervals. Fig. 1 presents an excerpt from a turn segmentation of a spoken dialog. In this paper we investigate the performance of three algorithms for turn boundary detection, the modulation spectrum analysis (MSA), Kullback-Leibler (KL) divergence for adjacent signal portions and Bayesian Information Criterion (BIC) for boundary detection. We propose an HMM-based combination of the event detection algorithms and show that their combination is outperforms each of them. The evaluation of the turn segmentation algorithms is carried out over the Switchboard corpus and assessed for a varying decision tolerance time window.

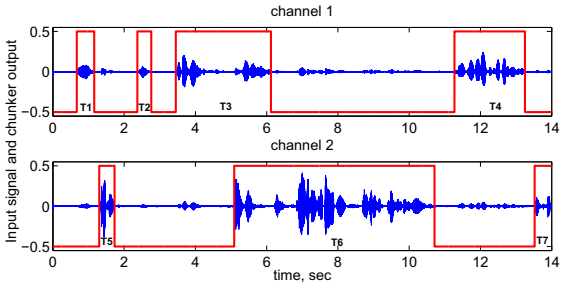

Figure 1: *Example of a dialog turn (T_i) segmentation. Turns "T_1" - "all right"; "T_2" - "uh"; "T_3" - "we've got a lot of them [laughter] too many of them"; "T_4" - "yeah well my husband's real good at using them"; "T_5" - "okay"; "T_6" - "oh yes [laughter] and i believe we all do and and it's just too easy to use"; "T_7" - "[laughter]".*

2. Algorithm description

Our speech segmentation algorithm is constructed as a combination of feature extraction techniques which feed a final decision generation process. It takes input from the various stages of a conventional MFCC feature computation procedure (i.e. signal power and 12 cepstral coefficients of the Mel-spectrum and their respective first and second derivatives at a rate of 100 frames per second) to save the computational power.

2.1. Modulation Spectrum Analyzer (MSA)

The modulation spectrum analysis (MSA) algorithm [5, 6] uses a stream of short-time Fourier transform (STFT) frames as its

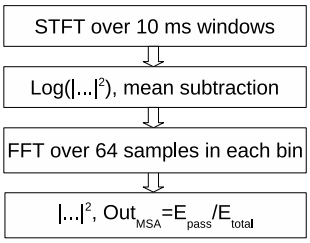

Figure 2: *MSA algorithm schematic diagram.* E_{pass}/E_{total} *- a ratio of modulation power in the range of interest to the total.*

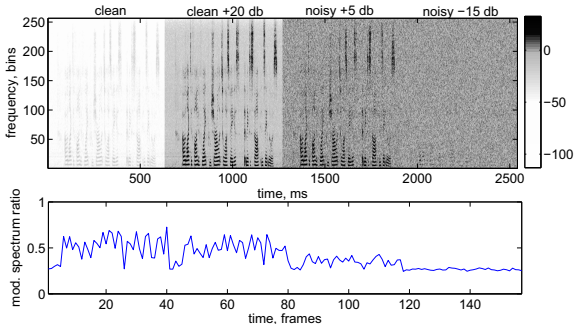

Figure 3: *MSA of speech examples. Four experiments: MSA of the original signal; MSA of the signal enhanced by 20dB; MSA of the signal enhanced by 20dB + Gaussian noise (SNR 5dB); MSA of the original signal + Gaussian noise (SNR -15dB).*

input (see Fig. 2). In the experiments reported in this paper the size of the STFT observation window was chosen to be 64 frames. Each of the spectral bins is considered a signal in time, for which a Fourier transform can be computed. This spectrum reflects how fast the energy of the respective frequency band is getting alternated through time.

The Fourier transform in the log-power domain is even more revealing. Let us consider a "source-channel" model of speech production, when the constant source signal (either a wide band noise or a harmonic complex) is convolved with a variable speech production channel. In the log-spectral domain the time-domain convolution becomes an addition. Thus, by taking a Fourier transform along the individual STFT log-spectral trajectories it is possible to isolate and characterize the dynamical properties of speech production apparatus. This representation of the signal is known as a modulation spectrum [6]. The final signal representation is three-dimensional, having frequency, modulation frequency and time as the axes.

Speech signals typically have a peak in modulation spectrum in the range between 1 and 10 Hz. This fact can be explained roughly through the "effective" rate of phoneme production being in the range of 1 to 10 times per second at maximum. We should note, however, that the rate of alternation of plosives is much higher than that of vowels and fricatives. In order to abstract away from the distribution across different spectral channels we sum modulation power spectral distributions to obtain one cumulative modulation power spectrum at any time instance. Then with the help of a simple filter in the modulation spectrum domain we estimate the power in 1 to 10 Hz modulation range. The final output of the modulation spectrum analyzer is a ratio of power in the modulation spectrum, contained in the range of interest to the total modulation power of the signal being analyzed E_{pass}/E_{total}.

The output of the MSA is approximately invariant to the incoming signal scale (see the first two signals on Fig.3). The output does not vary much with the input signal being enhanced 20 dB. Thus, the dynamic range of the MSA algorithm is reasonably large compared with the dynamic range of the pulse-coded modulation (PCM) signal representation.

Additive noise attenuates the distinctive modulation spectral signature of the signal and eventually makes it undetectable at approximately -15 dB SNR (see the last three signals on Fig.3). Although the moderate level of Gaussian additive noise is not a realistic scenario in telephone conversations, its effect is similar to the effect of quantization noise in PCM representation. We have observed in our experiments that signals, whose amplitude is comparable with quantization step, get lower MSA scores because of the quantization noise. In the situation of mutual across-channel interference MSA does not mark faint echoes from the other channel as "speech". However, for situations with higher level of across-channel interference a special source separation technique must be employed.

The typical value of MSA output for ideal clean speech is approx. 0.4-0.5, while typical stationary noise gets 0.2. It is not a power in the particular sub-band, but rather the effective rate of signal alternation through frequency subbands which allows us to discriminate between speech/non-speech events. Our experiments have revealed that such nonstationary signals, like tractor noise, helicopter noise, windshield wipers do not lead to MSA output being in the "speech" range. However, certain music genres, bird songs are still confusable with speech for the current version of MSA algorithm.

2.2. Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) algorithm is aimed at the detection of significant changes in the sample statistics. The BIC-feature is being extracted in two phases. First, the observation interval is split around a hypothesized statistics changing point T_0, thus producing a buffer X for the signal values immediately preceding a hypothesized boundary, a buffer Y for the data, which immediately follows the hypothesized boundary and the buffer Z which is a union of X and Y. In the performed experiments the size of buffers X and Y was chosen to be 64 frames of the MFCC feature set.

A simple multivariate Gaussian is estimated from the data in the buffers. The respective parameter sets are being denoted by θ_x, θ_y and θ_z. The second stage includes estimation of the probabilities to encounter particular observation with a given initial assumption of the model. The final BIC distance is computed with the formula (1) [7]. This formula reflects a competition between two hypotheses: one is that model θ_z is better to describe the whole set of data (both buffers X and Y), the other is that a combination of different models θ_x (for buffer X) and θ_y (for buffer Y) is better:

$$D_{BIC} = \sum_{i=1}^{N_x} \log \frac{p(x_i|\theta_x)}{p(x_i|\theta_z)} + \sum_{i=1}^{N_y} \log \frac{p(y_i|\theta_y)}{p(y_i|\theta_z)} - \\ - \left(n + \frac{n(n+1)}{2}\right) \frac{\lambda \log N_Z}{2}. \quad (1)$$

where $N_Z = N_X + N_Y$ is a total number of observations in the analysis window; n is the dimensionality of the feature vector; λ is a data-dependent parameter and needs to be tuned to achieve the best performance.

2.3. Kullback-Leibler Divergence (KLD)

The same set of buffers X, Y and Z can also be used in computing Kullback-Leibler Divergence (KLD) as a measure of dis-

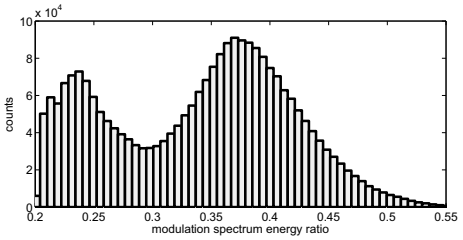

Figure 4: *Histogram of MSA readings.*

tribution similarity. As in the case of BIC the data in the three buffers is modelled as a single multivariate Gaussian distribution. The KLD from a true underlying Gaussian distribution D_0 to a modelling Gaussian distribution D_1 is defined as follows:

$$d_{KL}(D_0||D_1) = \frac{1}{2}(log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + tr\left(\Sigma_1^{-1}\Sigma_0\right) + \\ + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - n), \quad (2)$$

where n is the feature vector dimensionality.

As KLD is not a symmetric distance (i.e. generally $d_{KL}(A||B) \neq d_{KL}(B||A)$) it must be decided which of the buffers gives a statistical estimate of the "true" underlying distribution of data and which is to be used to compute a "model" distribution in question. Our experiments revealed little difference between the possible ways to use different buffers in KLD computation, but the sum D_{KL} of distances $d_{KL}(N_X||N_Z)$ and $d_{KL}(N_Y||N_Z)$ has lead to slightly better results.

2.4. Decision Combination Algorithm

The decision of the algorithm is synthesized with a two-state ("silence" and "speech") ergodic HMM via the process of Viterbi-decoding as a state sequence most probably corresponding to the observed acoustic evidence.

Emission probability distributions in this HMM are modelled by Gaussian fitting to the statistical evidence of MSA responses on conversational speech data. We have used a separate set of 15 dialog sides from the Switchboard database [8]. The emission penalty P_E is being computed from the emission probability by taking a negative logarithm.

Fig. 4 presents a histogram of the MSA output. The distribution is very close to a bi-modal sum of Gaussians. According to the properties of the MSA algorithm, each of these Gaussians corresponds to a particular state of the decision generating HMM. Thus, essentially the emission model has emerged from the unlabeled data in an unsupervised manner.

Transition penalties P_T in the HMM are not constant but rather governed by either the D_{BIC} or D_{KL}, which are linearly scaled to have the same mean and equal dynamic ranges. This allows to reduce the penalty associated with an HMM transition whenever the output of the BIC or KLD algorithm reports a significant mismatch in the statistics of the future and past buffers. Scaling parameters are being found as ones that maximize recognition performance on the set of training data.

To measure the relative impact of individual algorithms on the generated decisions "information-less" substitutions were employed. In the "MSA only" case there was no signal to generate a variable transition penalty, thus a constant value of $\tilde{P}_T$ was used instead. In the cases of the BIC and KLD being exclusively used, an emission penalty $\tilde{P}_E$ was linearly proportional to a logarithm of the state duration. In these conditions both HMM states have equal properties and there is no way to predict the type of transition. In this sense the task in which the KLD and BIC are being evaluated is simpler than a task for the MSA and combinations of the algorithms.

In order to facilitate an on-line processing of the audio data with the spoken turn segmenter the procedure of continuous backtracking has been implemented. At each frame the HMM transition histories are being compared and an update of their common history is reported.

3. Performance Evaluation

The algorithms are evaluated on the data taken from the Switchboard corpus [8] (LDC97S62). The test set contained 100 dialogs recorded as duplex communication, which yields 200 speaker channels. Recordings from the database were taken at random with no attention to the audio quality. This was done to estimate an expected performance level of the algorithm in realistic "loosely controlled" environments. The signal quality annotation which was done after selection, has revealed that the test set has the following major problems. There are recordings with both channels containing speech from both speakers mixed in different proportions. Recordings contain a large amount of high intensity clicks. There are recordings with non-stationary background interference, e.g. music, baby cries. The reference manual segmentation for the experiment was taken from the latest release[1] of the Switchboard re-segmentation project [9].

Table 1 presents a comparative study of the algorithms and their combinations. They are also compared to the performance of the reference implementation of the VAD of the standard G.729b vocoder [3]. The tolerance interval Δt is chosen to be 1 second, which appears to be appropriate for the spoken turn segmentation task. The operating points of all systems in the table (except G.729b VAD) were chosen in order to maximize F1-score as measured on the training set.

Table 1: *Performance of different predictors. "Δt" – tolerance interval, "CR" – total number of correct recognitions, "FR" – total number of false rejections, "FA" – total number of false acceptances, "F1" – F1-measure*

Detector Type	Δt, sec	CR	FR	FA	F1
MSA + KLD	**1.00**	**19134**	**7655**	**10904**	**0.6734**
MSA + BIC	1.00	19580	7209	12252	0.6680
MSA alone	1.00	19665	7124	12816	0.6636
MSA alone	0.75	18954	7835	13751	0.6418
MSA alone	0.50	18004	8785	14785	0.6044
MSA alone	0.25	16324	10465	16471	0.5479
MSA alone	0.10	12898	13891	19898	0.4329
BIC alone	1.00	24785	2004	55991	0.4608
KLD alone	1.00	16989	9800	37839	0.4163
G.729b	1.00	26671	118	264903	0.1675

All of the reviewed algorithms perform vastly better in spoken turn detection in comparison with the VAD of G.729b. Having a very small false rejection rate, the latter produces many false boundary detections. This fact comes as a result of the vocoder VAD being designed for a different task. As mentioned above, in voice coding the main purpose of the VAD is to isolate silence intervals during which, the communication bandwidth can be spared. Being very sensitive to the energy of the underlying signal, this VAD produces far too finely grained and not very robust to noise sequence of the silence and speech segments. We have observed that the false acceptance rate for this VAD does not vary much depending on the true segment label. This observation suggests that a postprocessing technique is not

[1] http://www.isip.piconepress.com/projects/switchboard/

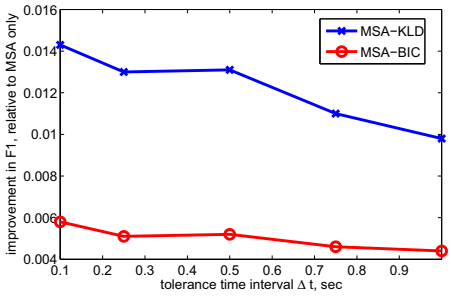

Figure 5: *Relative performance improvement wrt "MSA alone" of the combined predictors ("MSA-KLD" and "MSA-BIC") with different tolerance intervals.*

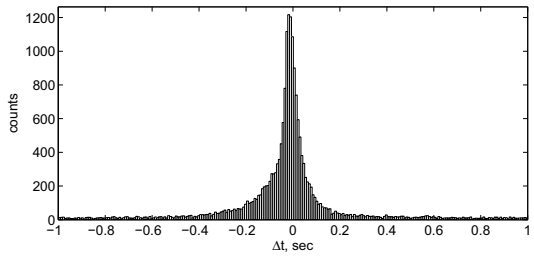

Figure 6: *Histogram of misalignment of the boundary prediction compared to the true position of the boundary.*

likely to help to recover spoken turn segmentation from the output stream of this VAD.

Both KLD and BIC used alone perform worse then the MSA. BIC appears to be a better predictor compared to KLD. This result can be explained in terms of a mismatch between the task and algorithm design principles.These algorithms are designed to detect an arbitrary change in the input signal statistics, this change should not necessarily be a transition between silence and speech. Originally the proposed application for these algorithms in speech science was in speaker diarization.

The analysis buffer duration used in these experiments is too short to gather the sample statistics reliably. That is why BIC and KLD algorithms produce "noisy" decisions with relatively high false acceptance rates. We have seen that an appropriate change of parameters can still bring FA and FR rates to a balance, but at the cost of a large increase of the FR rate and, thus, reduction of the F1-measure.

The combination of MSA and BIC or KLD produce slightly better results than MSA alone. Despite its lower computational complexity the MSA-KLD hybrid performs better than MSA-BIC. This observation is counterintuitive as the KLD alone is the worst performer.

We have also assessed performance of the systems when the tolerance Δt reduces down to hundred milliseconds. As it can be seen from both Table 1 and Fig. 5 the degradation does not dramatically accelerate even in the case of the tightest tolerance window. The hybrids are progressively more precise compared to the case of "MSA alone" with the tolerance window getting shorter. The KLD and BIC are algorithms in time domain and can provide finer boundary positioning compared to the frequency-domain MSA.

Fig. 6 depicts a histogram of misalignment of the boundary prediction compared to the true position of the boundary. The measurement is done for the best performing MSA-KLD hybrid. The distribution is close to a Gaussian, but its variance towards the positive values of the misalignment is smaller compared to that towards the negative values. The peak of the distribution occurs around the misalignment value of -0.01 sec, which is equal to the repetition rate of the observation frames in time (either STFT or the final feature vectors). Thus we can conclude that the bias of the predictor is close to its minimum.

4. Conclusions

A comparison of the different speech detection and segmentation methods has revealed that an HMM combination of MSA and KLD performs at the best level in the task of detection of spoken turns in natural telephone dialogs. The proposed method works directly from the speech recording and unlike existing methods of lexical segmentation does not rely on the existence of a transcription. In this task it outperforms traditional VAD algorithms as it aims at minimization of the false decision rate rather then maximization of correctly labeled data ratio.

The algorithm is to be used to segment speech data during live conversations for subsequent recognition of the complete individual spoken turns. The systematic delay of the segmentation procedure, while depending on the complexity of the task of speech detection in given conditions, has a lower bound of around half a second. This is deemed to be acceptable for employment in realtime automated dialog analysis systems.

5. Acknowledgements

This work was partially supported by the European Commission Marie Curie Excellence Grant for the ADAMACH project (contract No. 022593).

6. References

[1] Liu Y., Shriberg E., Stolcke A., Hillard D., Ostendorf M., and Harper M., "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Trans. on Speech and Audio. Process.*, vol. 14, no. 5, pp. 1526–1540, 2006.

[2] J.M. Górriz, J. Ramírez, and C.G. Puntonet, "New Advances in Voice Activity Detection Using HOS and Optimization Strategies," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds., p. 460. I-Tech, 2007.

[3] ITU, "A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation v.70," *ITU-T Recommendation G.729-Annex B*, 1996.

[4] ETSI, "Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Trafc Channels," *ETSI EN 301 708 Recommendation*, 1999.

[5] J.-H. Bach, B. Kollmeier, and J. Anemüller, "Modulation–Based Detection of Speech in Real Background Noise: Generalization to Novel Background Classes," in *Proc. of Int. Conf. on Acoust. Speech and Signal Processing (ICASSP)*, March 2010, pp. 41–44.

[6] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of Speech from Nonspeech Based on Multiscale Spectrotemporal Modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 920–930, May 2006.

[7] M. Kotti, V. Moschou, and C. Kotropoulos, "Review. Speaker Segmentation and Clustering," *Signal Processing*, vol. 88, pp. 1091–1124, 2008.

[8] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Proc. of Int. Conf. on Acoust. Speech and Signal Processing (ICASSP)*, March 1992, pp. 517–520.

[9] J. Hamaker, N. Deshmukh, A. Ganapathiraju, and J. Picone, "Resegmentation and Transcription of the SWITCHBOARD Corpus," in *Proc. of Speech Transcription Workshop*, September 1998.