# Speech/Nonspeech Segmentation in Web Videos

*Ananya Misra*

Google, New York, NY, USA

amisra@google.com

## Abstract

Speech transcription of web videos requires first detecting segments with transcribable speech. We refer to this as segmentation. Commonly used segmentation techniques are inadequate for domains such as YouTube, where videos may have a large variety of background and recording conditions. In this work, we investigate alternative audio features and a discriminative classifier, which together yield a lower frame error rate (25.3%) on YouTube videos compared to the commonly used Gaussian mixture models trained on cepstral features (30.6%). The alternative audio features perform particularly well in noisy conditions.

**Index Terms**: segmentation, speech detection, voice activity detection, video

## 1. Introduction

Speech/nonspeech segmentation in the presence of background noise is an important first step for automatic speech recognition, and has received attention from various sources. Common techniques for segmentation include frame energy-based methods and Gaussian mixture models (GMMs) trained on Mel-frequency cepstral coefficients (MFCCs), with a hidden Markov model (HMM) or other smoothing module.

In the broadcast news domain, the use of HMMs with GMM emissions of MFCCs has been well studied [1]. Further investigations include hierarchical segmentation integrated with factor analysis [2], and energy-based methods combined with conditional random fields (CRF), which also use the MFCC GMM log likelihoods as a feature [3]. Additional features such as PLPs [1] and Chroma [2] have been used in those studies.

Voice activity detection also builds on similar techniques. Kalman filter noise models have been added to combine noise suppression or adaptation with voice activity detection [4, 5]. Other features studied for voice activity detection include a periodic-aperiodic ratio [4], autocorrelation-based voicing features [6] and relative energy measurements in different frequency bands [7]. Fundamental frequency and lower-level signal analysis have also been used with MFCC GMM log likelihoods and mixture posteriors as features to a CRF [8].

Content analysis and speech/music discrimination efforts present other features of interest [9, 10, 11]. Several of the features investigated in the current work, such as energy, zero crossing and flux-based features, as well as line spectral pairs, have previously been used for hierarchical speech/nonspeech segmentation of TV news, movie clips, and internet audio clips [12].

While many of these ideas might aid speech/nonspeech segmentation in uploaded web videos, none has been particularly evaluated for such a task. However, segmentation is especially important for uploaded videos, many of which have little or no speech. Recent work focusing on web videos observed that amateur videos are particularly hard to segment due to ample noise and recording artifacts [13]; in fact, it omitted such "homebrew" videos altogether due to the challenges presented. Broadcast news, which appears to be the closest well-explored area, does not present the same challenges. In this work, we focus on novel classification approaches and features, informed by other domains, that are more robust to such conditions. We evaluate these on a mixed collection of videos selected without human curation.

In the rest of the paper, we define the novel features (Section 2) and classifiers (Section 3) explored, and describe experimental results on a mixed web video data set (Section 4).

## 2. Features

The baseline (`MFCC`) feature set consists of 13 MFCCs along with their first and second derivatives, with the cepstrum normalized over each waveform. The following alternative features (`Alt`) are tested against the baseline.

**Low short-time energy ratio:** The ratio of frames with a short-time energy below $x = 0.5$ times the average in a larger window around a given frame [12]. While frame energy is commonly used for speech detection [3, 7], this feature considers the characteristics of a larger surrounding window; it is expected to be higher in windows with speech than music or noise, assuming speech has more pauses and thus more energy fluctuations [12].

**High zero-crossing rate ratio:** The ratio of frames with zero-crossing rate above $x = 1.5$ times the average in a larger window around a given frame. As zero-crossing rate is related to pitch, this is also expected to be higher in speech, which has alternating voiced and unvoiced sections [12].

**Line spectral pairs (LSP):** These are transformations of linear predictive coding (LPC) coefficients that lie on the unit circle and thus correspond to isolated frequencies. They are expected to distinguish between speech, noisy speech and music. Previous work [12] builds a speech codebook from covariance matrices of the LSPs within speech windows, and labels a test window by thresholding the distance of its covariance matrix to this codebook. This work uses raw LSPs.

**Spectral flux:** This measures the change in spectral amplitudes between consecutive frames. Speech is expected to alternate between periods of change and stability while music has a more constant rate of change [9, 11].

**Spectral centroid:** The center of mass of the spectrum, this is expected to give different results for voiced and unvoiced speech as well as music, since spectral energy is concentrated in different regions for each [9, 11].

**Spectral rolloff:** The frequency point at which the energy at lower frequencies is equal to $x = 90\%$ of the energy at higher frequencies. This is expected to distinguish between unvoiced speech, which has more energy in the higher bands, and voiced

speech or music [9, 11].

**Ratio of magnitudes in speech band:** The ratio of the sum of magnitudes in frequency bands that typically contain speech, compared to the sum of magnitudes over the whole spectrum. This is expected to be higher for voiced speech than other classes, and is similar to a relative energy by frequency band metric [7].

**Top peaks:** The top $n = 5$ peaks in the magnitude spectrum, defined by frequency and the corresponding frame-normalized magnitude. These are expected to represent the dominant frequencies in the frame, and thus to be different for voiced speech, music and noise.

**Ratio of magnitudes under top peaks:** This measures the ratio of the sum of magnitudes of the top $n$ peaks to the sum of magnitudes over the whole spectrum. Voiced speech, music and other harmonic sounds are expected to have a higher proportion of energy concentrated in the top spectral peaks.

# 3. Classifiers

Two frame-level classifiers are compared, with and without a smoothing layer. The study focuses on a binary classification of speech or nonspeech, with arbitrary recording and background conditions.

## 3.1. Frame-level Classifers

### *3.1.1. Gaussian Mixture Models (GMM)*

For an $N$-dimensional observed feature vector $o$, and an $M$-component GMM, the likelihood of observation $o$ at state $s_i$ is:

$$P(o \mid s_i) = \sum_{j=1}^{M} \omega_{ij} \mathcal{N}(o \mid \mu_{ij}, \Sigma_{ij}) \quad (1)$$

where $\omega_{ij}$, $\mu_{ij}$ and $\Sigma_{ij}$ are respectively the mixture weight, mean and diagonal covariance for the $j^{th}$ component of state $s_i$, and $\mathcal{N}$ is a normal distribution. GMMs with a maximum of $64$ components each were trained for speech and nonspeech, using maximum likelihood (ML). Maximum mutual information (MMI) training on MFCC GMMs also yielded similar performance. Overall, this model has $6M$ (or $384$) parameters and a size of $6MN$.

### *3.1.2. Maximum Entropy Classifier (Maxent)*

A maximum entropy classifier is a discriminative model with the greatest entropy from among those that approximately satisfy a given set of constraints [14]. This study uses a conditional maximum entropy model. Given the input space of feature vectors $O$ and output space of labels $S$ (in this case, binary), the conditional probability of a label $s_i$ given an observation $o$ is defined as:

$$P_{\mathbf{w}}(s_i \mid o) = \frac{e^{\mathbf{w} \cdot \phi(o, s_i)}}{\sum_{s \in S} e^{\mathbf{w} \cdot \phi(o, s)}} \quad (2)$$

In this case, $\phi$ is an identity mapping from $O \times S$ to itself. The weight vector $\mathbf{w} \in O \times S$ is learned by maximizing the log probabilities $P_{\mathbf{w}}(s_j \mid o_j)$ over the training data. The output predicted by the model for a given observation is the label that maximizes this conditional probability [14]. Note that for binary output, $\phi(o, s)$ can be replaced by a mapping on the input alone, with a single weight associated with each element of the input vector. Hence the size and parameters of the model are equivalent to the feature dimensionality ($N$).

## 3.2. Sequence-level Smoothing

Smoothing here refers to de-noising frame-by-frame classifications to yield longer contiguous speech and nonspeech segments. It is commonly achieved with an ergodic hidden Markov model (HMM), in this case with "nonspeech" and "speech" states and equal transition cost in either direction. There is no cost for remaining in the same state, which creates a smoothing effect by discouraging state changes. The transition cost is selected empirically; section 4.4 suggests that this value has a large effect on the results and should ideally be learned from the data or otherwise automatically tuned.

The emission probabilities, $b_j(o)$, represent the probability of observation $o$ given state $s_j$. For the GMM classifier, these are equal to the GMM emissions (Equation 1).

Because the Maxent classifier is discriminative, it does not immediately provide probabilities of the form $P(o \mid s_j)$. However, it is possible to estimate a prior distribution of states from the training data and set $b_j(o) = P(s_j \mid o)/P(s_j)$. In this case, given that the training data had close to equal amounts of each state, the probabilities given by the Maxent classifier were directly used.

# 4. Experiments

## 4.1. Data Sets

This work primarily uses an anonymized data set of 95 hours of YouTube video, divided into 90 hours for training and 5 hours for testing. Videos were sampled from YouTube in an automated way based on usage statistics, with more popular videos having a greater chance of selection to support evaluation on seemingly important videos. Video segments were manually defined and labeled as silence, music, noise, speech, or combinations of these labels. Segments had a granularity of 5 seconds or more for tractability and in light of naturally occurring pauses in informal speech. Some statistics are presented in Table 1. Note that the classes are not mutually exclusive; for instance, "speech with noise" and "speech with music" both include segments containing speech, noise and music.

| Class | Training | Test |
|---|---|---|
| Clean speech | 14.2h | 46m |
| Silence | 1.6h | 6m |
| Speech with noise | 19.6h | 84m |
| Noise without speech | 11.2h | 52m |
| Speech with music | 16.7h | 63m |
| Music without speech | 39.8h | 134m |

Table 1: *Data set statistics showing the durations of training (hours) and test (minutes) data for selected classes.*

For both training and evaluation, speech samples included speech with all backgrounds, while nonspeech samples included everything without speech. This supported the goal of detecting speech in arbitrary backgrounds by reducing the mismatch between training and test conditions. However, similarly labeled data can easily be leveraged to train and evaluate more specific models.

## 4.2. Evaluation Metrics

The models are evaluated using the following metrics. All but the equal error rate are based on the durations of correct, missed and false alarm speech at a fixed operating point.

- **False alarm rate** (FA): The percentage of nonspeech that is classified as speech.
- **Miss rate** (Miss): The percentage of speech that is classified as nonspeech.
- **Equal error rate** (EER): The point on the ROC curve (which charts miss and false alarm rates over varying acceptance thresholds) where the miss rate equals the false alarm rate. This evaluates frame-by-frame error before any smoothing.
- **Segmentation error** (SegErr) = (missed speech + false alarm speech) / total reference time. This is based on the NIST speaker diarization error rate without individual speaker information. It is also similar to the metric used by Castan *et. al.* [2], but looks at only the speech error scaled by the total reference duration.

### 4.3. Frame-by-frame results

Frame-by-frame errors for the `MFCC` and `Alt` feature sets (see Section 2) are described in Table 2, for both the Maxent and GMM classifiers. For each experiment, an overall EER is given as well as the EER in specific background conditions. For example, the "Music" column shows the EER for segments that include music, corresponding to the last two rows of Table 1. Results for the clean background condition are highly variable, partly due to the small amount of clean data, especially silence, in both the training and test sets (see the first two rows of Table 1); these are still included for completeness.

| Experiment | Overall | Music | Noise | Clean |
|---|---|---|---|---|
| Maxent | | | | |
| MFCC | 40.8 | 43.6 | 46.2 | 52.5 |
| Alt | 25.3 | 29.5 | 32.2 | 4.4 |
| GMM | | | | |
| MFCC | 30.6 | 33.0 | 41.1 | 7.8 |
| Alt | 33.8 | 36.8 | 38.3 | 9.3 |
| Augmented MFCC+Maxent | | | | |
| GMM scores | 29.2 | 30.4 | 42.6 | 16.3 |
| Cross-products | 31.7 | 36.1 | 37.7 | 17.0 |

Table 2: *Equal error rate (EER)% from ROCs based on frame-by-frame classifications, before smoothing. EER on backgrounds containing music, noise or neither are also shown. The final three rows indicate Maxent experiments with the MFCC features augmented either by scores from the MFCC GMM or by taking cross-products of the MFCC features.*

Note that in general, the lowest error rates appear for the `Alt` feature set and the Maxent classifier. This suggests that the features and classifier complement each other in a way suitable for frame-by-frame speech/nonspeech classification. With the GMM classifier, `Alt` yields a lower EER for data with background noise, suggesting that the new features contribute some amount of noise robustness in their own right, even without a new classifier. This may follow from the fact that a number of the alternative features look at signal characteristics that distinguish between speech and pure noise.

The poor performance of `MFCC` with the Maxent classifier suggests that MFCCs may be better suited to a generative model. In particular, the covariance between individual features in a frame can be better modeled by a mixture of Gaussians. Since the MFCC coefficients together estimate a spectral envelope, their covariance may be especially important. Augmenting the `MFCC` features with speech and nonspeech scores from the MFCC GMM greatly improved `MFCC+Maxent` performance. Alternatively, expanding the `MFCC` feature set to include cross-products of MFCC coefficients also led to improved performance. These results are depicted in the final section of Table 2.

### 4.4. Post-smoothing error

Applying HMM-based smoothing to the frame-by-frame classification resulted in a neutralization of the gains observed in Section 4.3. Results for Maxent and GMM with the various feature sets are described in Table 3.

| Classifier | Features | SegErr | FA | Miss |
|---|---|---|---|---|
| Maxent | MFCC | 46.8 | 9.5 | 82.3 |
| | Alt | 22.2 | 22.3 | 22.2 |
| GMM | MFCC | 19.6 | 24.1 | 15.2 |
| | Alt | 22.3 | 22.6 | 22.0 |

Table 3: *Results for each feature set and classifier with HMM smoothing. SegErr shows the overall segmentation error%, while FA and Miss show the false alarm rate and miss rate respectively, as percentages.*

The difference in per-frame and post-smoothing results suggests that the smoothing technique plays a large role in determining the final accuracy of the system. This is natural, as per-frame results are expected to be noisy; not only does the data set have a reference granularity of seconds, but some unvoiced speech, such as fricatives, may have noise-like characteristics.

Further experiments confirm that smoothing parameters are also important. Figure 1 shows results for varying HMM transition weights for two interesting models: `Alt+Maxent` and `MFCC+GMM`. It reveals that while each of these sees some gain from smoothing, the GMM model gets the largest relative gain. It is interesting to note that for noisy backgrounds, `Alt+Maxent` outperforms `MFCC+GMM` even after smoothing.

It is also plausible that because the HMM-based smoothing is set up to fit a generative context (see Section 3.2), a different smoothing technique may be ideal for the discriminative Maxent classifier.

### 4.5. Analysis

We analyzed roughly 400 features based on the feature sets discussed above. These included all the raw features used in the previous experiments as well as first and second derivatives of the `Alt` feature set. The square of each of these features was also considered. Finally, the scores of speech and nonspeech MFCC GMMs were also evaluated as features.

Features were evaluated on a per-frame basis without smoothing, using a conditional mutual information criterion [15]. This selects the next "best" feature one at a time, such that it is the one least subsumed by any of the already selected features individually. Note that only pairwise dependencies are considered.

This analysis found the 10 most important features to be:

1. Low short time energy ratio squared (LSTER$^2$)
2. Nonspeech GMM score
3. Freq0 (frequency of highest spectral peak) squared
4. Speech GMM score
5. MFCC 12 squared

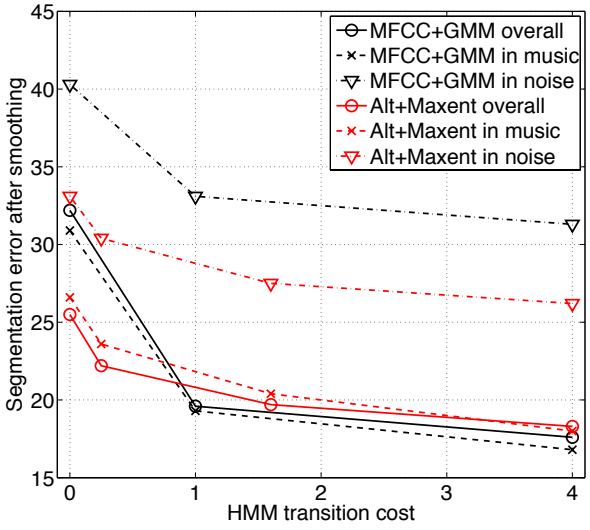


Figure 1: *Segmentation error (SegErr)% for Alt+Maxent and MFCC+GMM, for different HMM transition costs and background conditions.*

6. High zero crossing rate ratio squared
7. First derivative of freq0, squared
8. Second derivative of freq0, squared
9. Energy between 85 and 255 Hz
10. First LSP coefficient

This suggests that both the MFCC GMM scores and the alternative features capture meaningful dimensions without being redundant. Note that the highest spectral peak as well as its first and second derivatives appear on this list. Since this peak can represent a crude pitch approximation, the way it changes across frames is also relevant. Note also that many of these features appear more useful when squared, suggesting that higher-order, kernel-based classifiers would be beneficial.

## 5. Conclusion

An alternative set of features and classifier were shown to provide better frame-by-frame accuracy for classifying speech versus nonspeech in web videos. These features also performed better than conventional MFCCs under noisy conditions, with both the GMM and Maxent classifiers (see Table 2). Conditional mutual information analysis on a superset of experimental features found metrics derived from several of the new features to be among the top differentiators for speech versus nonspeech (see Section 4.5).

While some of this gain is neutralized upon smoothing, the smoothing technique explored in this paper is heuristic, especially when combined with the Maxent classifier. A more principled approach is needed, perhaps in the form of CRFs or voting schemes, or by estimating real transition probabilities in an HMM or maximum entropy Markov model. Similarly, while Section 4.5 suggests that the `MFCC` and `Alt` features complement each other, intelligent methods to combine the results from different feature sets and/or classifiers need to be explored. Such explorations are outside the scope of this paper, although existing work suggests possible directions [8, 16].

While opportunities for further research abound, this paper establishes that features and classifiers beyond the conventional system can benefit speech/nonspeech segmentation, especially in arbitrary web videos with little control over noise and recording conditions.

## 6. Acknowledgements

Many thanks to colleagues for their help, comments and insights; in particular: Chris Alberti, Dan Bikel, Erik McDermott, Gideon Mann, Hank Liao, Konstantinos Katsiapis, Michiel Bacchiani, Olivier Siohan, Patrick Nguyen, Shankar Kumar and Thomas Deselaers.

## 7. References

[1] T. Hain and P. C. Woodland, "Segmentation and classification of broadcast news audio," in *Proc. of ICSLP*, 1998, pp. 2727–2730.

[2] D. Castán, C. Vaquero, A. Ortega, D. Martínez, J. Villalba, and E. Lleida, "Hierarchical audio segmentation with hmm and factor analysis in broadcast news domain," in *Proc. of Interspeech*, 2011, pp. 421–424.

[3] C. Gao, G. Saikumar, S. Khanwalkar, A. Herscovici, A. Kumar, A. Srivastava, and P. Natarajan, "Online speech activity detection in broadcast news," in *Proc. of Interspeech*, 2011, pp. 2637–2640.

[4] M. Fujimoto, K. Ishizuka, and T. Nakatani, "Study of integration of statistical model-based voice activity detection and noise suppression," in *Proc. of Interspeech*, 2008, pp. 2008–2011.

[5] R. J. Weiss and T. Kristjansson, "DySANA: Dynamic speech and noise adaptation for voice activity detection," in *Proc. of Interspeech*, 2008, pp. 127–130.

[6] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Proc. of Interspeech*, 2005, pp. 369–372.

[7] T. Dekens and W. Verhelst, "On noise robust voice activity detection," in *Proc. of Interspeech*, 2011, pp. 2649–2652.

[8] A. Saito, Y. Nankaku, A. Lee, and K. Tokuda, "Voice activity detection based on conditional random fields using multiple features," in *Proc. of Interspeech*, 2010, pp. 2086–2089.

[9] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. of ICASSP*, 1997, pp. 1331–1334.

[10] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *Proc. of ICASSP*, 1999, pp. 149–152.

[11] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. on Speech and Audio Proc.*, vol. 10, no. 5, pp. 293–302, July 2002.

[12] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. on Speech and Audio Proc.*, vol. 10, no. 7, pp. 504–516, October 2002.

[13] P. Clement, T. Bazillon, and C. Fredouille, "Speaker diarization of heterogeneous web video files: A preliminary study," in *Proc. of ICASSP*, 2011, pp. 4432–4435.

[14] G. Mann, R. McDonald, M. Mohri, N. Silberman, and D. Walker, "Efficient large-scale distributed training of conditional maximum entropy models," *In Advances in Neural Information Processing Systems*, vol. 22, pp. 1231–1239, 2009.

[15] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, pp. 1531–1535, December 2004.

[16] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, March 1998.