Frontiers in Emerging Computer Science and Information Technology (Open Access)

Volume 02, Issue 08, August 2025,
Publish Date: 01-08-2025
PageNo.01-06

Robust Speech Segmentation Through Non-Speech Interval Suppression

Dr. Michael J. Thompson
Centre for Speech Technology Research, University of Edinburgh, United Kingdom

Dr. Yoon-Seok Kim
Department of Electrical and Computer Engineering, Seoul National University, South Korea

ABSTRACT

Speech segmentation, the process of dividing continuous speech into meaningful units such as words, syllables, or phonemes, is a fundamental precursor for numerous speeches processing applications, including automatic speech recognition (ASR), speaker diarization, and language acquisition studies. Traditional segmentation methods often struggle with robustness in challenging acoustic environments, particularly in the presence of varied background noise and silent intervals. This article proposes and outlines a novel framework for robust speech segmentation based on the principle of "Non-Speech Interval Suppression" (NSIS). This method primarily focuses on accurately identifying and eliminating silent or non-speech regions to precisely delineate the boundaries of spoken content. By leveraging advanced Voice Activity Detection (VAD) techniques and integrating them within a structured segmentation pipeline, NSIS aims to enhance segmentation accuracy, especially in noisy conditions. The proposed framework's potential to yield cleaner speech segments is discussed, leading to improved performance in downstream speech technologies and offering a more robust foundation for speech analysis.

KEYWORDS: Speech segmentation, non-speech interval suppression, speech processing, voice activity detection, audio signal processing, robust segmentation.

INTRODUCTION

Speech is the primary mode of human communication, and its automatic processing forms the bedrock of numerous technological advancements, from virtual assistants to real-time translation services. A critical initial step in almost any speech processing pipeline is speech segmentation, the task of dividing a continuous audio stream into smaller, meaningful units. These units can range from broad segments like speech vs. non-speech, to finer distinctions such as words, syllables, or phonemes [1]. Accurate segmentation is paramount for the success of subsequent processing stages, including feature extraction, acoustic modeling, and ultimately, automatic speech recognition (ASR) [1, 7]. For instance, a precise definition of speech boundaries can significantly improve the performance of ASR systems by focusing computational resources solely on relevant acoustic information [13].

Historically, researchers have explored various approaches to speech segmentation. Early attempts focused on basic acoustic properties like energy and zero-crossing rate to delineate speech events [2]. More sophisticated methods have emerged, including those based on hierarchical classification of audio data for archiving and retrieval [4], and speaker-independent phoneme classification in continuous speech [5]. The study of how humans, particularly infants, segment fluent speech using cues like phonotactics [3] has also provided insights into potential computational models. Furthermore, the discovery of novel word-like units from utterances, even in artificial language studies, underscores the importance of segmentation in language acquisition models [6].

Despite these advancements, robust speech segmentation remains a significant challenge, particularly in real-world

scenarios characterized by varying levels of background noise, reverberation, and speaker variability. Unwanted "black areas," or non-speech intervals (e.g., silence, environmental noise, or music), often contaminate speech recordings, leading to inaccurate segmentation, increased computational load, and degraded performance in downstream applications. The presence of these non-speech components can mislead algorithms designed to identify linguistic units, thereby hindering the overall efficacy of speech processing systems. Consequently, there is a compelling need for segmentation methodologies that can effectively identify and suppress these non-speech intervals, thereby enhancing the precision and reliability of speech segmentation in diverse and challenging acoustic environments. This article aims to propose a conceptual framework centered on this "Non-Speech Interval Suppression" (NSIS) principle to address these persistent challenges.

METHODS

The proposed framework for Robust Speech Segmentation through Non-Speech Interval Suppression (NSIS) fundamentally relies on the accurate identification and exclusion of silent or background noise regions within an audio stream. This method leverages advanced Voice Activity Detection (VAD) techniques as its core mechanism, transforming raw audio into cleaner segments that are more amenable to further linguistic processing. The methodology can be broken down into several interconnected steps:

Audio Signal Pre-processing:

Framing: The continuous audio signal is first divided into short, fixed-duration frames (e.g., 10-30 ms), typically with significant overlap to ensure smooth transitions and capture transient speech events.

Windowing: Each frame is then multiplied by a window function (e.g., Hamming window) to minimize spectral leakage at the frame boundaries.

Feature Extraction for Voice Activity Detection:

To differentiate between speech and non-speech, various acoustic features are extracted from each frame. While basic features like short-time energy and zero-crossing rate are fundamental, more advanced features are crucial for robust VAD in noisy environments.

Spectral Features: Mel-Frequency Cepstral Coefficients (MFCCs) are widely used for their robustness to channel variations and their representation of the spectral envelope of speech.

Perceptual Features: Features derived from psychoacoustic models can better reflect how humans perceive sound, aiding in discriminating speech from noise.

Prosodic Features: Pitch (fundamental frequency) and its variations, along with energy contours, provide strong indicators of voiced speech segments.

Voice Activity Detection (VAD) – The "Blocking" Mechanism:

This is the critical step where non-speech intervals are identified and marked for suppression. Several robust VAD algorithms, often drawing from various research contributions, can be employed:

Model-Based VAD: Approaches utilizing hidden Markov models (HMMs) are effective for robust voicing and speech detection, modeling speech and noise characteristics over time [11].

Rule-Based VAD: Algorithms employing fuzzy rules can effectively classify voiced/unvoiced segments, providing a flexible way to handle ambiguous acoustic events [9]. These rules often combine multiple features to make decisions.

Hybrid Feature VAD: Combining various acoustic features (e.g., spectral, temporal, prosodic) with a network classifier can lead to highly accurate voiced-unvoiced-silence classification [10]. This approach leverages the complementary strengths of different feature sets.

Array Signal Processing in Wavelet Domain: For multi-microphone setups, techniques involving array signal processing in the wavelet domain can significantly enhance voice activity detection by spatially filtering noise and leveraging the time-frequency properties of speech signals [8]. This is particularly useful in environments with multiple noise sources.

The output of the VAD module is typically a binary decision for each frame: speech or non-speech.

Non-Speech Interval Suppression and Boundary Refinement:

Once VAD identifies non-speech frames (the "black areas"), these intervals are conceptually "blocked" or suppressed. This means that subsequent segmentation processes will primarily operate on the identified speech-active regions.

Initial Segmentation: Based on the VAD output, the audio stream is segmented into contiguous speech blocks.

Fine-Grained Boundary Detection: Within these speech blocks, more granular segmentation (e.g., into syllables or phonemes) can occur. This fine-tuning often involves:

Energy and Zero-Crossing Rate: Re-evaluating these features at the edges of speech blocks to precisely locate syllable or phoneme boundaries, as described in early automatic segmentation methods [2].

Acoustic Change Points: Identifying points where acoustic properties significantly change, indicating potential linguistic boundaries [12].

Phonotactic Constraints: For phoneme or word-level segmentation, integrating linguistic knowledge (phonotactics) can guide the segmentation process, similar to how infants segment speech [3].

Blind Speech Segmentation: Methods that automatically segment speech without requiring explicit linguistic knowledge can also be integrated, particularly for initial or broad segmentation tasks [14].

Automatic Phonetic Transcription: For more advanced applications, the segmented speech can then be fed into systems for automatic phonetic transcription, which inherently performs a fine-grained segmentation [15].

Post-Processing and Evaluation:

Merge/Split Operations: Small gaps within speech regions or short isolated speech bursts can be handled through adaptive merging or splitting logic to create more coherent segments.

Evaluation: The performance of the NSIS framework is evaluated using metrics such as detection error rates for speech activity, and accuracy of subsequent word or syllable recognition tasks, comparing it against established benchmarks and traditional segmentation methods [1, 12, 13].

By systematically suppressing non-speech intervals as an initial and integral step, the NSIS framework provides a cleaner, more focused input for downstream speech analysis, promising enhanced robustness and accuracy across a wide range of applications and challenging acoustic conditions.

RESULTS

The implementation of the Robust Speech Segmentation through Non-Speech Interval Suppression (NSIS) framework is hypothesized to yield significant improvements in speech processing pipelines, particularly in environments characterized by noise and variability. The primary results anticipated from the application of this methodology are:

Enhanced Accuracy of Speech/Non-Speech Classification: By employing sophisticated Voice Activity Detection (VAD) techniques, including those based on hybrid features and network classifiers [10], fuzzy rules [9], and HMMs [11], the NSIS framework is expected to achieve a higher precision in distinguishing active speech from silence, background noise, or other non-linguistic sounds. This translates to a lower rate of false positives (misclassifying noise as speech) and false negatives (missing actual speech segments). For example, leveraging array signal processing in the wavelet domain [8] for VAD can dramatically improve performance in multi-source noisy environments, leading to cleaner initial segmentation.

Improved Robustness in Challenging Acoustic **Environments:** One of the main benefits of explicit non-speech interval suppression is the enhanced resilience of the segmentation process to varying noise levels and types. By effectively "blocking out" contaminated regions, the subsequent linguistic segmentation algorithms operate on higher-fidelity speech data, making them less susceptible to degradation caused by environmental interference. This is crucial for real-world applications where recordings are rarely pristine.

More Precise Speech Segment Boundaries: With non-speech intervals reliably identified and suppressed, the framework enables the delineation of speech unit boundaries (e.g., words, syllables, phonemes) with greater accuracy. This is particularly important for applications requiring fine-grained temporal alignment, such as phonetics research or high-quality concatenative speech synthesis. The initial broad segmentation from VAD allows for focused, fine-tuned boundary detection within active speech regions using techniques that might otherwise be confused by noise [12, 14].

Reduced Computational Load for Downstream Tasks: By pre-filtering non-speech content, the NSIS framework ensures that subsequent computationally intensive processes, such as acoustic modeling for ASR or feature extraction for speaker identification, are applied only to relevant speech data. This leads to more efficient resource utilization and faster processing times for a complete speech recognition architecture [1].

Higher Performance in Downstream Speech Processing Applications:

Automatic Speech Recognition (ASR): Cleaner input segments from NSIS are expected to lead to lower word error rates (WER) in ASR systems, as the acoustic models are not attempting to recognize non-speech noise [13].

Speaker Diarization and Identification: By accurately segmenting speech by speaker, and removing non-speech interruptions, the framework can improve the accuracy of speaker identification and diarization systems.

Language Acquisition Studies: For studies focusing on how infants discover word-like units [6] or utilize phonotactic cues [3], robust segmentation ensures that the analyzed input truly represents relevant linguistic information, leading to more reliable research findings.

Audio Archiving and Retrieval: Hierarchical classification of audio data for archiving and retrieval [4] benefits immensely from accurate speech/non-speech and speaker segmentation, allowing for more precise indexing and search functionalities.

In summary, the NSIS framework is designed to provide a more refined and robust foundation for all subsequent speech processing tasks. By focusing on the precise identification and suppression of non-speech content, it is anticipated to significantly elevate the quality of speech segments, leading to measurable improvements across a spectrum of speech technology applications.

DISCUSSION

The proposed Robust Speech Segmentation through Non-Speech Interval Suppression (NSIS) framework offers a compelling solution to a long-standing challenge in speech processing: achieving accurate and robust segmentation in diverse and often noisy acoustic environments. By placing emphasis on the precise identification and suppression of non-speech intervals, the framework leverages advanced Voice Activity Detection (VAD) as a foundational "blocking" mechanism, thereby providing cleaner and more reliable input for subsequent, finer-grained segmentation tasks.

A key strength of this approach lies in its proactive handling of noise and silence. Unlike methods that attempt to segment speech directly within noisy conditions, NSIS first purifies the signal by delineating active speech regions. This modularity allows for the integration of highly specialized VAD techniques, such as those employing fuzzy rules [9], hybrid features with network classifiers [10], or advanced HMMs [11], each contributing to the robust detection of speech. The ability to incorporate array signal processing in the wavelet domain [8] further strengthens the framework's performance in challenging multi-source noise scenarios, which are common in real-world applications. This initial cleansing of the audio stream sets a strong foundation for subsequent linguistic segmentation, enhancing the overall reliability.

The framework's focus on non-speech suppression directly addresses a critical pain point in current speech processing. Inaccurate segmentation due to noise can lead to degraded performance in Automatic Speech Recognition (ASR), misidentification in speaker recognition, and flawed data in phonetic analysis [1]. By ensuring that only active speech content is passed on for further processing, NSIS can significantly reduce computation on irrelevant data, potentially leading to faster and more efficient systems,

which is a key objective for a new speech recognition architecture [1]. The potential for improved word error rates in ASR and more precise acoustic-phonetic analysis aligns well with the fundamental principles of speech recognition [13].

Despite its promising advantages, the NSIS framework also presents certain limitations and areas for future exploration. The performance of the entire framework is highly dependent on the accuracy of the initial VAD module. A VAD system that frequently misclassifies speech as noise (false negatives) or noise as speech (false positives) will inevitably propagate errors through the segmentation pipeline. Achieving near-perfect VAD in highly complex and dynamic noise environments (e.g., overlapping speech, sudden loud noises) remains an active research area. While various VAD methods are cited [8, 9, 10, 11], the optimal choice and tuning will be context-dependent.

Furthermore, while NSIS effectively handles the speech/non-speech boundary, the fine-grained linguistic segmentation within the identified speech regions still relies on sophisticated algorithms. Whether segmenting into syllables [2, 7] or phonemes [5, 15], integrating linguistic knowledge and robust acoustic feature analysis is crucial. The framework provides a cleaner canvas, but the artistry of detailed segmentation still requires advanced techniques. Blind speech segmentation methods [14], which do not rely on linguistic knowledge, could be a starting point, but often require refinement for specific applications.

Future research could focus on several key areas. First, exploring the application of deep learning models for end-to-end NSIS, where VAD and segmentation are jointly optimized, could potentially yield further improvements in robustness and accuracy. Second, investigating adaptive VAD techniques that can learn and adjust to changing noise characteristics in real-time would enhance the framework's utility in highly dynamic environments. Third, the quantification of performance gains across a wider range of languages and acoustic conditions, including challenging scenarios found in actual speech corpora [1], would provide stronger empirical validation for the NSIS approach. Finally, integrating the framework with specific applications like speech emotion recognition or forensic voice analysis could reveal new challenges and opportunities for refinement.

CONCLUSION

The Robust Speech Segmentation through Non-Speech Interval Suppression framework offers a powerful and intuitive approach to enhancing the precision and robustness of speech segmentation. By systematically eliminating contaminating non-speech intervals, it provides a purer signal for linguistic analysis, paving the way for more accurate, efficient, and reliable speech processing systems across a myriad of applications.

REFERENCES

1. Okko Rasanen. (2007). *Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture*. M.Sc Thesis, Department of Electrical and Communications Engineering, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Espoo, November.
2. Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America*, 58(4), 880–883.
3. Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78, 91–121.
4. Zhang, T., & Kuo, C.-C. J. (1999). Hierarchical classification of audio data for archiving and retrieving. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 6, 3001–3004.
5. Antal, M. (2004). Speaker independent phoneme classification in continuous speech. *Studia Universitatis Babeş-Bolyai, Informatica*, 49(2).
6. Dahan, D., & Brent, M. R. (1999). On the discovery of novel word-like units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, 128, 165–185.

7. Thangarajan, R., & Natarajan, A. M. (2008). Syllable-based continuous speech recognition for Tamil. *South Asian Language Review*, 18(1).
8. Hioka, Y., & Namada, N. (2003). Voice activity detection with array signal processing in the wavelet domain. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 86(11), 2802–2811.
9. Beritelli, F., & Casale, S. (1997). Robust voiced/unvoiced classification using fuzzy rules. In *1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings* (pp. 5–6).
10. Qi, Y., & Hunt, B. (1993). Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier. *IEEE Transactions on Speech and Audio Processing*, 1(2), 250–255.
11. Basu, S. (2003). A linked-HMM model for robust voicing and speech detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*.
12. Kvale, K. (1993). *Segmentation and Labeling of Speech*. PhD Dissertation, The Norwegian Institute of Technology.
13. Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall.
14. Sharma, M., & Mammone, R. (1996). Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP '96)*, Vol. 2, 1237–1240.
15. Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. In *Proceedings of the ICPhS*, San Francisco, 607–610.
16. Shapiro, L. G., & Stockman, G. C. (Year not provided). *Computer Vision*. [Incomplete reference—please provide full details if available.]