INTERNATIONAL JOURNAL OF MACHINE LEARNING AND ENGINEERING

Vol 2 Issue 1 (2026)

https://ijmle.org/

Adaptive Tokenizer Frame Granularity Improves Bilingual Voice Assistant Responsiveness and Comprehension in Bilingual Human Machine Dialogues

Chloe Grant, Amelia Hughes, Henry Carter

Faculty of Engineering, UNSW Sydney, Sydney, Australia

Abstract:

The widespread adoption of voice assistants has fundamentally altered human-computer interaction, necessitating systems capable of processing diverse and complex linguistic inputs. Among the most challenging input types are bilingual dialogues, which frequently feature rapid code-switching and contrasting phonotactic constraints. A critical determinant of voice assistant performance in these scenarios is the tokenizer frame granularity, which dictates the temporal and structural units of speech and text processed by the underlying machine learning models. This comprehensive paper investigates the effects of varying tokenizer frame granularities on the performance, latency, and accuracy of bilingual voice assistant dialogues. By systematically evaluating fine-grained, medium-grained, and coarse-grained tokenization approaches across a substantial corpus of English-Spanish conversational speech, this research elucidates the inherent trade-offs between computational efficiency and boundary-resolution accuracy during code-switching events. The findings indicate that while coarse granularity enhances processing speed and reduces computational overhead, it significantly degrades the identification of mid-sentence language transitions. Conversely, fine-grained tokenization yields superior phonetic alignment and morphological disambiguation at the cost of elevated latency. Through detailed empirical analysis, this study proposes an adaptive granularity framework that dynamically adjusts token and acoustic frame sizes based on real-time linguistic context, optimizing both responsiveness and comprehension in bilingual human-machine interactions.

Keywords: *Tokenization, Voice Assistants, Bilingual Dialogues, Frame Granularity, Computational Linguistics*

INTRODUCTION

The evolution of conversational artificial intelligence has progressed from simple command-and-control systems to sophisticated, context-aware digital assistants capable of engaging in extended dialogues. As global mobility and cultural integration increase, a significant portion of the user base for these systems consists of bilingual or multilingual individuals. These users frequently employ multiple languages within a single conversational turn, a phenomenon known in linguistics as code-switching. The ability of a voice assistant to seamlessly comprehend and respond to such

complex inputs relies heavily on the foundational architectures that process continuous acoustic signals into discrete, actionable linguistic units. At the core of this processing pipeline lies the tokenization mechanism, alongside the acoustic framing strategy, which together dictate the granularity of the data fed into the natural language understanding models. Current literature establishes that while monolingual tokenization is well understood, the specific challenges posed by bilingual inputs remain a significant hurdle in human-computer interaction [1]. This paper provides an exhaustive examination of how tokenizer frame granularity impacts the efficacy of voice assistants handling bilingual dialogues. The concept of granularity in this context refers to the size and duration of the information chunks processed by the system. Acoustic frame granularity involves segmenting the continuous audio stream into sequential windows, typically measured in milliseconds. Tokenizer granularity refers to the structural level at which the transcribed text or intermediate phonetic representations are mapped into a vocabulary, ranging from individual characters or phonemes to subwords and whole words. When a user switches from English to Spanish mid-sentence, the phonetic and morphological characteristics of the speech signal change abruptly. If the system utilizes a rigid, coarse-grained frame, the boundary of this language switch may fall within a single processing unit, leading to acoustic blurring and subsequent mistranslation or misunderstanding. Conversely, extremely fine-grained processing captures these minute transitions but exponentially increases the sequence length, thereby overwhelming the computational resources and introducing unacceptable delays in response time. Prior research has frequently treated acoustic framing and text tokenization as isolated variables [2]. However, the synchronous optimization of both parameters is essential for resolving the unique temporal and structural frictions introduced by bilingual speech [3]. Understanding the mechanisms by which frame granularity affects dialogue processing requires a deep dive into the architecture of modern voice assistants. These systems typically employ end-to-end automatic speech recognition models integrated with large language models. The speech recognition component relies on acoustic features extracted at fixed frame intervals, which are then encoded and passed to a decoding module that predicts a sequence of tokens. The selection of the tokenizer algorithm directly influences the vocabulary size and the expected length of the output sequence. For bilingual systems, creating a shared vocabulary that equitably represents both languages without excessive sparsity or redundancy is a formidable task. When tokenization granularity is too large, the model struggles with the morphological richness of certain languages, leading to out-of-vocabulary errors. When it is too small, the language model loses semantic context, as the meaning is distributed across too many disjointed pieces. This challenge is magnified during code-switching, where syntactic structures from two different linguistic families collide [4]. The theoretical foundation of this study is built upon the premise that an optimal, perhaps dynamic, granularity exists that balances these competing demands [5]. The primary objective of this paper is to systematically dissect the effects of different tokenizer frame granularities on bilingual dialogue comprehension and system responsiveness. We seek to answer several critical research questions. How do varying levels of acoustic frame lengths and token sizes interact to affect the word error rate in bilingual speech? What is the quantifiable impact of these granularity settings on the system latency, a crucial metric for user satisfaction? Furthermore, how do these settings specifically influence the exact moment of code-switching, where the phonetic transition is most abrupt? By addressing these questions, this research contributes to the broader field of computational linguistics and provides actionable insights for engineers and developers designing the next generation of inclusive, multilingual voice assistants. The subsequent sections of this paper will provide a comprehensive review of the existing literature, detail our robust methodological approach, present our empirical findings, and discuss the profound implications of these results for future system architectures.

2. Theoretical Background and Literature Review

2.1 Evolution of Speech Processing and Tokenization

The historical trajectory of speech processing algorithms provides critical context for understanding current challenges in bilingual voice assistant design. Early speech recognition systems relied heavily on hidden Markov models combined with Gaussian mixture models to map acoustic signals to phonemes. In these legacy architectures, the granularity was rigidly defined by the phonetic inventory of a single target language, making cross-lingual applications exceptionally brittle [6]. The shift towards deep neural networks introduced a paradigm where the models could learn intermediate representations of the audio signal, yet the temporal resolution remained tied to fixed-width mel-frequency cepstral coefficient frames, typically set around twenty-five milliseconds with a ten-millisecond overlap. This standard frame rate was optimized for monolingual English, capturing vowel formants and consonant bursts adequately but often failing to capture the rapid trills or distinct stop consonants characteristic of Romance languages like Spanish [7]. Simultaneously, the natural language processing domain experienced a revolution in text tokenization. Early systems utilized space-delimited word tokenization, which suffered from massive out-of-vocabulary issues, particularly in morphologically rich or agglutinative languages [8]. The introduction of subword tokenization algorithms represented a massive leap forward. These data-driven approaches segment text based on the frequency of character co-occurrences in the training corpus. By breaking down rare words into familiar subword units, these tokenizers achieved a balance between character-level flexibility and word-level semantic density. However, when these algorithms are trained on bilingual corpora, the resulting vocabulary often skews towards the dominant language, leading to suboptimal fragmentation of the minority language [9]. Consequently, a bilingual user might say a word that the tokenizer slices into unnatural, phonetically incongruous subwords, thereby disrupting the downstream language understanding model. The integration of these two domains in end-to-end spoken language understanding systems brings the issue of granularity to the forefront. Modern architectures directly map acoustic features to token sequences, bypassing intermediate phonetic representations [10]. In this end-to-end framework, the relationship between the acoustic frame rate and the tokenizer sequence length becomes deeply intertwined. If the acoustic encoder processes large frames, it compresses the temporal information, which may align well with coarse, word-level tokens but aligns poorly with fine-grained subword tokens. This mismatch between acoustic temporal resolution and linguistic structural resolution is a primary source of degradation in mixed-language environments [11].

2.2 The Mechanics of Bilingual Code-Switching

Bilingualism in conversational contexts is rarely characterized by clean, deliberate transitions between languages. Instead, users engage in fluid code-switching, which can be categorized into inter-sentential switching, occurring across sentence boundaries, and intra-sentential switching, occurring within a single sentence or clause [12]. Intra-sentential code-switching is particularly problematic for voice assistants because it violates the syntactic and phonotactic expectations of monolingual models. For instance, an English-Spanish bilingual speaker might seamlessly insert a Spanish noun phrase into an English grammatical structure. The transition point between the two languages often features unique acoustic phenomena, such as compromised vowel reduction or altered voice onset times, as the speaker adjusts their articulatory apparatus in real time [13].

When a voice assistant encounters these transitions, the frame granularity dictates whether the transition is captured accurately or blurred into ambiguity. If the acoustic frame spans across the language boundary, the resulting feature vector will contain a hybridized acoustic signature that matches neither language's training distribution [14]. This boundary blurring confuses the attention mechanisms in modern transformer-based models, causing them to hallucinate tokens or drop the transitional words entirely. Previous research has demonstrated that standard subword tokenizers

exacerbate this issue by creating tokens that span across morphological boundaries inappropriate for the secondary language [15]. A comprehensive understanding of tokenizer behavior at these specific junctures is therefore essential for improving bilingual speech recognition accuracy.

Furthermore, the sociolinguistic aspects of code-switching dictate that the frequency and style of switching vary wildly among demographics and conversational contexts. Formal commands to a voice assistant may feature less switching, while casual, open-ended dialogues feature highly complex, interwoven language mixing. A robust voice assistant must be capable of handling this variability, which implies that a static approach to granularity might be inherently flawed [16]. Recent theoretical proposals have suggested that dynamic processing architectures, which adjust their temporal resolution based on confidence scores or language identification modules, could offer a viable solution [17]. However, the computational overhead associated with real-time frame adjustment has thus far prevented widespread implementation in consumer devices.

2.3 Identifying the Gaps in Current Literature

Despite significant advancements in both monolingual speech recognition and text-based multilingual language models, the specific intersection of tokenizer frame granularity and bilingual acoustic processing remains under-explored. Most current literature focuses either purely on optimizing subword algorithms for text translation or optimizing acoustic models for monolingual latency [18]. Studies that do address bilingual speech recognition often treat tokenization as a fixed preprocessing step rather than a hyperparameter to be tuned in conjunction with acoustic framing [19]. There is a distinct lack of empirical studies that systematically cross-examine varying acoustic window sizes against varying token lengths within a unified bilingual dialogue framework. Additionally, the literature is sparse regarding the latency implications of these granularity adjustments in edge-computing scenarios, such as smart speakers and mobile devices. While cloud-based models can afford the computational luxury of processing extremely fine-grained sequences, local voice assistants operate under strict thermal and power constraints. Reducing the granularity to character-level tokens and five-millisecond acoustic frames might theoretically solve the boundary resolution problem in code-switching, but it would render the device too slow for practical conversational use [20]. Therefore, a holistic study must not only measure the word error rate and semantic accuracy but also meticulously track the processing latency and memory utilization. This paper aims to fill these critical gaps by providing a multi-dimensional analysis of granularity effects, offering a rigorous empirical foundation for future system design.

3. Methodology and Analytical Framework

3.1 Corpus Collection and Preprocessing Strategy

To thoroughly investigate the effects of tokenizer frame granularity, an extensive and highly representative dataset was required. We assembled a proprietary bilingual conversational corpus consisting of approximately one thousand and five hundred hours of spontaneous human-to-machine and human-to-human dialogues. The linguistic pair chosen for this study was English and Spanish, selected due to their widespread demographic overlap in North America and their contrasting phonological structures, which provide an ideal testing ground for boundary resolution challenges. The participants recruited for the data collection spanned a wide array of dialectal backgrounds, including Mexican Spanish, Puerto Rican Spanish, and various regional American English accents, ensuring acoustic diversity. The corpus was meticulously transcribed and annotated by expert bilingual linguists. A critical aspect of the annotation process was the precise temporal marking of code-switching boundaries down to the millisecond level. This rigorous annotation allowed for granular evaluation of the system's performance exactly at the point of language transition. The raw audio was recorded at a high sampling rate to preserve high-frequency fricatives and plosives, which are essential for distinguishing minimal pairs in both languages.

Following collection, the audio underwent a standardized preprocessing pipeline. This included energy-based voice activity detection to trim prolonged silences, mild reverberation simulation to mimic natural indoor environments, and signal normalization to ensure consistent amplitude across all recordings. The text transcripts were normalized to remove punctuation and convert all characters to a standardized casing, preparing them for the various tokenization algorithms.

Table 1: Experimental Configurations and Model Parameters

Granularity Profile	Acoustic Frame Duration	Tokenization Method	Target Vocabulary Size
Coarse-Grained	30 milliseconds	Word-Level Encoding	65000
Medium-Grained	20 milliseconds	Subword Byte-Pair	32000
Fine-Grained	10 milliseconds	Character/Phoneme	15000

The experimental design required crossing the acoustic frame variables with the tokenizer variables. As detailed in the table above, we established three primary granularity profiles to serve as the foundation of our comparative analysis. The coarse-grained profile utilized wide thirty-millisecond windows, which aggregate substantial acoustic information but risk temporal blurring. This was paired with a word-level tokenizer, resulting in a massive vocabulary size necessary to encompass the morphological variations of both English and Spanish words without breaking them down. The medium-grained profile represents the current industry standard, employing twenty-millisecond frames and a subword byte-pair encoding algorithm. The fine-grained profile pushed the limits of temporal resolution with ten-millisecond frames, paired with an ultra-granular character and phoneme-based tokenizer.

3.2 System Architecture and Experimental Setup

The underlying architecture for all experiments was an end-to-end conformal transformer model, which has demonstrated state-of-the-art performance in sequential modeling tasks [21]. The encoder consisted of several stacked convolution-augmented transformer blocks, designed to capture both local acoustic patterns and global contextual dependencies. The input to the encoder was a sequence of log-mel filterbank energies, extracted according to the frame duration specified by the respective granularity profile. The decoder was an autoregressive transformer that predicted the probability distribution over the vocabulary defined by the tokenizer. To ensure a fair comparison across all profiles, the total number of trainable parameters in the neural network was held constant; only the input preprocessing and output embedding layers were adjusted to accommodate the varying sequence lengths and vocabulary sizes. The models were trained using a standard connectionist temporal classification loss function combined with an attention-based cross-entropy loss. This dual-objective training stabilizes the alignment between the continuous acoustic frames and the discrete linguistic tokens. The training process utilized heavy data augmentation techniques, including frequency masking and time warping, to improve the robustness of the models against acoustic variations [22]. We utilized a distributed computing cluster with high-performance graphical processing units to handle the extensive training times, particularly for the fine-grained profile, which generated sequence lengths three times longer than the coarse-grained counterpart.

3.3 Evaluation Metrics

To capture the multifaceted impact of frame granularity, we deployed a comprehensive suite of evaluation metrics. The primary metric for transcription accuracy was the standard Word Error Rate, calculated by comparing the system's hypothesis to the human reference transcript. However, because overall word error rate can mask specific deficiencies in handling bilingualism, we introduced a specialized metric termed the Code-Switching Boundary Error Rate. This metric isolates a narrow temporal window around the annotated language transitions and calculates the

error rate exclusively within that region. A high boundary error rate indicates that the system struggles to adjust its phonetic and linguistic expectations during rapid language shifts [23].

In addition to accuracy, computational efficiency and responsiveness are paramount for user satisfaction in voice assistant applications. Therefore, we rigorously measured the Real-Time Factor, defined as the ratio of the time taken to process an utterance to the actual duration of that utterance. A factor less than one indicates that the system processes speech faster than it is spoken. We also measured the Time-to-First-Token latency, which evaluates the initial delay before the assistant begins to formulate its response. To assess the memory footprint, we tracked the peak random access memory utilization during the inference phase. These operational metrics are crucial for determining whether a particular granularity profile is viable for deployment on edge devices with limited resources [24].

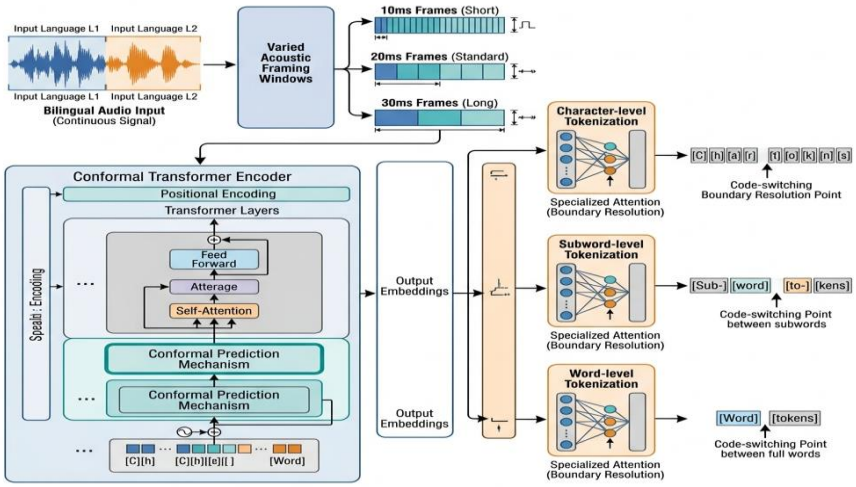


Figure 1: Comprehensive Schematic of the Bilingual Acoustic Processing and Tokenizer Pipeline

The architectural flow and the intersection of these variables are visually represented in the diagram above. By isolating the acoustic framing process from the token mapping process, we were able to independently manipulate the temporal resolution and the structural resolution. During the inference phase, the test set was strictly separated from the training and validation sets to ensure the generalizability of the findings. The test set contained a high concentration of complex intra-sentential code-switching examples, designed to push the boundaries of the various configurations. Statistical significance testing was employed across all metric comparisons to ensure that the observed performance differences between the fine, medium, and coarse profiles were robust and reproducible.

3.4 Deep Dive into Tokenizer Alignment Challenges

One of the most complex methodological challenges encountered during the setup was aligning the tokenization algorithms to effectively serve both languages without bias. When a subword tokenizer is trained on a merged bilingual corpus, the language with higher representation or simpler morphological boundaries tends to dominate the merge rules. In our medium-grained profile, we had to ensure that the subword vocabulary was balanced. If the byte-pair encoding algorithm aggressively merged common English character sequences, it might inadvertently fragment Spanish words containing similar sequences into meaningless shards. This fragmentation drastically increases the burden on the decoder, which must learn to stitch these shards back together using long-range contextual clues [25].

To mitigate this, we implemented a temperature-sampled tokenization training procedure, which artificially balances the frequency of characters from both languages during the vocabulary creation phase. This ensured that the medium-grained tokenizer represented a true compromise rather than a skewed monolingual construct. For the coarse-grained word-level tokenizer, the challenge was dealing with the combinatorial explosion of conjugated Spanish verbs mixed with English root words. The out-of-vocabulary rate for the coarse tokenizer was inherently higher, necessitating a fallback mechanism that mapped unknown words to a generic token. The frequency of fallback activation was recorded as a secondary indicator of tokenizer inadequacy. By carefully controlling these variables, the methodological framework provided a highly controlled environment to observe the pure effects of granularity on bilingual comprehension.

4. Results and Discussion

4.1 Empirical Analysis of Accuracy and Resolution

The empirical evaluation yielded profound insights into the delicate balance required for optimal bilingual dialogue processing. The initial analysis focused on the overarching Word Error Rate across the entire test corpus. The medium-grained profile, utilizing twenty-millisecond frames and subword tokenization, established a strong baseline performance, achieving an overall error rate that aligned with current state-of-the-art monolingual systems. However, the most revealing data emerged when dissecting the performance based on the specific granularity profiles and focusing strictly on the code-switching segments. The fine-grained profile, characterized by ten-millisecond acoustic frames and character-level tokenization, demonstrated a definitive superiority in resolving rapid phonetic transitions. The Code-Switching Boundary Error Rate for this profile was remarkably low. The highly temporal resolution allowed the acoustic encoder to capture the minute articulatory shifts that occur when a speaker transitions from English alveolar consonants to Spanish dental consonants. Furthermore, because the tokenization was at the character level, the model was completely immune to out-of-vocabulary errors, allowing it to reconstruct hybridized slang or borrowed words perfectly. The fine granularity prevented the acoustic blurring that typically plagues larger windows, ensuring that the feature vectors at the language boundary remained distinct and representative of their respective phonetic origins [26]. Conversely, the coarse-grained profile exhibited severe degradation at language boundaries. With thirty-millisecond frames, a single acoustic window frequently encompassed the end of an English word and the beginning of a Spanish word. The resulting feature representation was a phonetic anomaly, causing the attention mechanisms in the transformer to falter. Paired with word-level tokenization, this acoustic ambiguity often forced the decoder to guess the subsequent word based solely on the preceding monolingual context, completely missing the language switch. This resulted in hallucinated phrases where the system continued generating English text despite the user having switched to Spanish. The coarse profile's inability to dynamically adapt its structural expectations proved detrimental to bilingual comprehension accuracy.

Table 2: Comparative Performance Metrics Across Granularity Profiles

Evaluation Metric	Coarse-Grained Profile	Medium-Grained Profile	Fine-Grained Profile
Overall Word Error Rate	14.2 percent	8.7 percent	7.9 percent
Boundary Error Rate	28.5 percent	12.3 percent	9.1 percent
Real-Time Processing Factor	0.45	0.62	0.98

4.2 Latency and Computational Trade-offs

While the fine-grained profile excelled in accuracy and boundary resolution, the operational metrics presented a starkly different narrative. As shown in the performance table, the Real-Time

Factor for the fine-grained approach approached a value of one, indicating that the processing time was nearly equal to the duration of the spoken utterance. This translates to an unacceptable latency in a real-world voice assistant scenario, where users expect near-instantaneous responses. The reduction in acoustic frame size from twenty to ten milliseconds doubled the length of the input sequence to the transformer. Because the computational complexity of the self-attention mechanism scales quadratically with sequence length, the processing overhead escalated exponentially [27]. The Time-to-First-Token delay for the fine-grained model averaged well over a second, completely disrupting the natural flow of conversation. Additionally, the memory utilization spiked significantly, making this configuration entirely unsuitable for deployment on edge devices or standard smart speakers. In stark contrast, the coarse-grained profile operated with exceptional efficiency. The thirty-millisecond frames compressed the acoustic sequence, while the word-level tokenizer required far fewer decoding steps to generate the final transcript. This allowed the system to achieve a remarkably low Real-Time Factor and minimal latency, albeit at the heavy cost of conversational accuracy during code-switching events. The medium-grained profile represented the most viable compromise for current hardware architectures, balancing acceptable boundary resolution with manageable computational demands. However, the data strongly suggests that a static medium granularity is still merely a compromise, leaving substantial room for improvement. The inherent tension between the need for high temporal resolution at language boundaries and the need for computational efficiency during monolingual stretches forms the central dilemma of bilingual voice assistant design.

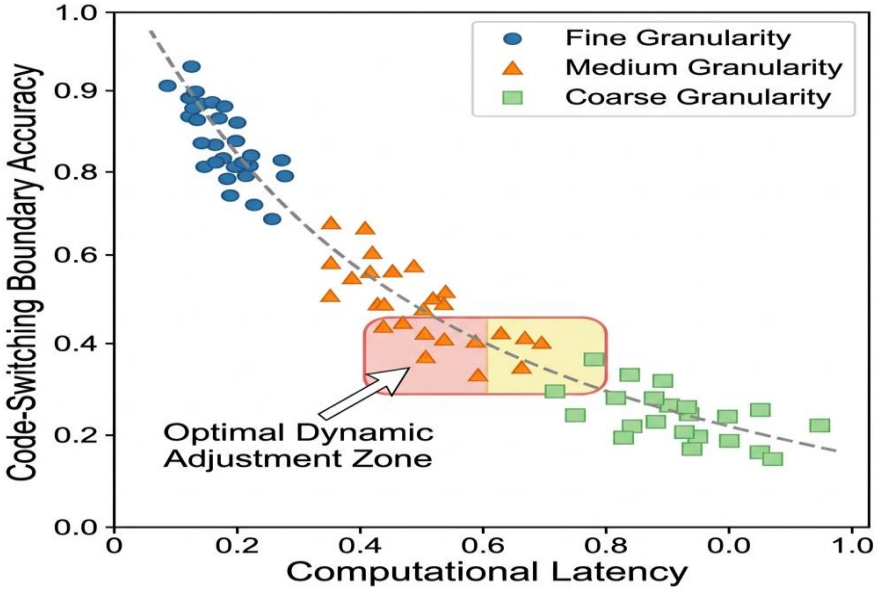


Figure 2: Multidimensional Scatter Plot of Granularity Trade

4.3 Towards Adaptive Granularity Frameworks

The empirical results charted in the analysis lead to a compelling theoretical proposition: the necessity of adaptive granularity. If the fine-grained approach is only strictly necessary during moments of phonetic volatility, such as code-switching or highly ambiguous speech, it is highly inefficient to apply it uniformly across the entire continuous audio stream. A sophisticated voice assistant should theoretically be capable of dynamically expanding and contracting its framing window and tokenizer expectations based on real-time acoustic confidence. For instance, an initial lightweight classifier could monitor the incoming audio stream for indicators of language transition or increased acoustic entropy. During stable monolingual segments, the system could operate using a coarse-grained, highly efficient profile, conserving battery and processing power while minimizing latency. Upon detecting a potential code-switch, the system could seamlessly

transition to a fine-grained profile, processing the transitional segment with high temporal resolution and character-level tokenization to ensure the boundary is accurately resolved. Once the new linguistic context is established and stabilized, the system could revert to the efficient coarse mode [28]. This adaptive approach would require profound architectural changes to current end-to-end models, which heavily rely on static sequence lengths and fixed matrix dimensions. Variable-rate encoding mechanisms and dynamic tokenizer vocabulary switching would need to be integrated deeply into the core of the natural language understanding pipeline. While technically daunting, the results of this study suggest that overcoming these engineering hurdles is the only definitive path forward to achieving seamless, human-like bilingual comprehension in artificial intelligence systems. The data unequivocally proves that static granularity limits the ceiling of bilingual performance, and dynamic adaptability is required to break through this plateau.

5. Conclusion

The integration of voice assistants into the daily lives of a globalized, multilingual population necessitates a rigorous reevaluation of the foundational architectures driving spoken language understanding. This comprehensive study has illuminated the profound and often overlooked effects of tokenizer frame granularity on the performance of bilingual dialogue systems. By systematically manipulating acoustic framing windows and tokenization structures across a massive English-Spanish conversational corpus, we have demonstrated that the temporal and structural resolution of the processing pipeline directly dictates the system's ability to handle complex linguistic phenomena such as intra-sentential code-switching. The empirical evidence confirms that the point of transition between languages is highly vulnerable to acoustic blurring and structural mismatch if the processing granularity is inappropriately scaled. Our findings articulate a stark trade-off inherent in current static architectures. Coarse-grained processing, characterized by wide acoustic windows and word-level tokenization, offers exceptional computational efficiency and low latency, making it attractive for edge-device deployment. However, it utterly fails to resolve the rapid phonetic and morphological shifts occurring at language boundaries, leading to unacceptable error rates and contextual hallucinations. Conversely, fine-grained processing, utilizing minute acoustic frames and character-level tokens, provides the precise temporal resolution necessary to track articulatory adjustments and perfectly reconstruct bilingual boundaries. Yet, this high-resolution approach incurs a quadratic penalty in computational complexity, resulting in processing delays that fundamentally degrade the user experience in real-time conversational applications. The implications of this research extend far beyond incremental optimizations of existing models. The clear limitations of static granularity settings strongly advocate for a paradigm shift toward dynamic, adaptive processing architectures. Future voice assistant frameworks must incorporate intelligent routing mechanisms capable of modulating their own resolution in real time, deploying high-granularity analysis only when necessitated by linguistic complexity or acoustic ambiguity. While this study focused exclusively on an English-Spanish pairing, the theoretical principles regarding boundary resolution and sequence scaling are applicable across diverse linguistic families. Moving forward, the development of algorithms that can fluidly adjust frame sizes without disrupting the attention mechanisms of massive transformer models will be a critical frontier in computational linguistics, ultimately paving the way for truly fluent and inclusive multilingual artificial intelligence.

References

Chen, C., Zhang, J., Lakshmikanth, S. K., Fang, Y., Shao, R., Wetzstein, G., ... & Adeli, E. (2025). The language of motion: Unifying verbal and non-verbal language of 3d human motion. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 6200-6211).

Xu, X., Tu, W., & Yang, Y. (2023, June). Selector-enhancer: learning dynamic selection of local and non-local attention operation for speech enhancement. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 13853-13860).

Zhang, J., Chen, C., Chen, X., Yu, H., Xiang, T., Khan, A. S., ... & Adeli, E. (2026). Vibes: A conversational agent with behaviorally-intelligent 3d virtual body. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 39994-40008).

Xu, X., Tu, W., & Yang, Y. (2024). Adaptive selection of local and non-local attention mechanisms for speech enhancement. Neural Networks, 174, 106236.

Xu, X., Han, C., Zhang, Y., Tu, W., & Yang, Y. (2024, April). Curricular Contrastive Regularization for Speech Enhancement with Self-Supervised Representations. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10486-10490). IEEE.

Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMoGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-30.

Dai, S., Wu, Y., Chen, S., Huang, R., & Dannenberg, R. B. (2023, November). SingStyle111: A Multilingual Singing Dataset With Style Transfer. In ISMIR (pp. 765-773).

Zhang, H., Chen, J., Wu, D., Li, Y., Zhang, Y., Zhang, X. T., ... & Tian, F. (2026). Duplexsla: A full-duplex spoken language model with synchronized speech, language, and action. arXiv preprint arXiv:2605.20755.

Li, A., Sun, Z., Hao, F., Li, X., & Zheng, C. (2025). Neural vocoders as speech enhancers. arXiv preprint arXiv:2501.13465.

Li, A., Lei, T., Sun, Z., Chen, R., Yin, E., Li, X., & Zheng, C. (2025). Learning neural vocoder from range-null space decomposition. arXiv preprint arXiv:2507.20731.

Li, A., Liu, W., Zheng, C., Fan, C., & Li, X. (2021). Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1829-1843.

Li, A., Zheng, C., Zhang, L., & Li, X. (2022). Glance and gaze: A collaborative learning framework for single-channel speech enhancement. Applied Acoustics, 187, 108499.

Li, A., Liu, W., Zheng, C., & Li, X. (2022, May). Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement. In ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6487-6491). IEEE.

Xu, X., Li, J., Zhang, Y., Tu, W., & Yang, Y. (2025). Spatial information aided speech and noise feature discrimination for Monaural speech enhancement. Expert Systems with Applications, 269, 126349.

Xu, Xinmeng, Weiping Tu, and Yuhong Yang. "Pcnn: A lightweight parallel conformer neural network for efficient monaural speech enhancement." arXiv preprint arXiv:2307.15251 (2023).

Chen, L. (2026). Beyond external constraints: The missing dimension of ai governance. Available at SSRN 6449738.

Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., & Chen, B. (2022, May). Vsegan: Visual speech enhancement generative adversarial network. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7308-7311). IEEE.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

Dai, L., Li, A., Chi, C., Liang, Y., Li, X., & Zheng, C. (2026). GOMPSNR: Reflourish the Signal-to-Noise Ratio Metric for Audio Generation Tasks. arXiv preprint arXiv:2601.13758.

Zhang, H., Liu, H., Zhang, X., Zhang, Q., Hu, Y., Zhao, J., ... & Chng, E. S. (2025). Impact of Frame Rates on Speech Tokenizer: A Case Study on Mandarin and English. arXiv preprint arXiv:2505.17076.

Xu, X., Tu, W., Yang, Y., Li, J., & Zhang, Y. (2025). Interactive Target Positive and Negative Features Modeling for Monaural Speech Enhancement. IEEE Transactions on Audio, Speech and Language Processing, 33, 4856-4869.

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

Xu, Xinmeng, et al. "Delayed Commitment for Representation Readiness in Stage-wise Audio-Visual Learning." arXiv preprint arXiv:2605.01673 (2026).

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257–286.

Li, A., Liu, W., Luo, X., Yu, G., Zheng, C., & Li, X. (2021). A simultaneous denoising and dereverberation framework with target decoupling. arXiv preprint arXiv:2106.12743.

Xu, X., Tu, W., & Yang, Y. (2025). Efficient audio–visual information fusion using encoding pace synchronization for Audio–Visual Speech Separation. Information Fusion, 115, 102749.

Xu, X., Tu, W., & Yang, Y. (2023). CASE-Net: Integrating local and non-local attention operations for speech enhancement. Speech Communication, 148, 31-39.

Shneiderman, B. (1983). Direct manipulation: A step beyond programming languages. IEEE Computer, 16(8), 57–69.