Acoustic Indicators of Join Quality in Concatenated Video Game Commentary

Luca Resti[13], *Amelia J. Gully*[2], *Michael McLoughlin*[3], *Gavin Kearney*[3], *Alena Denisova*[1]

[1]Department of Computer Science, University of York, United Kingdom
[2]Department of Language and Linguistic Science, University of York, United Kingdom
[3]AudioLab, School of Physics, Engineering and Technology, University of York, United Kingdom

{luca.resti, amelia.gully, michael.mcloughlin, gavin.kearney, alena.denisova}
@york.ac.uk

1. Introduction

In sports video games, commentary is sometimes generated by concatenating pre-recorded speech segments to dynamically narrate in-game events. While enabling reactive dialogue, this can introduce perceptual artifacts at join points, such as signal and prosodic discontinuities. Identifying these issues typically relies on manual quality control by audio designers, highlighting the need for data-driven approaches to support the review process. Our dataset consists of 18741 recordings of a professional sports commentator. Utterances are either names (teams or players) or actions (e.g., "passes the ball"). Full utterances are formed by randomly concatenating semantically compatible units on consonants, as per industrial practice. We analyzed 1,500 concatenated and 3,000 non-concatenated sports commentary samples using OpenSMILE's *ComParE_2016* feature set (6,373 acoustic features) [1]. This analysis lays the foundation for subjective testing and the development of a predictive model to assist the quality assurance process by identifying perceptually problematic joins.

2. Methodology

After standardizing the extracted features, we applied PCA followed by t-SNE and UMAP (as shown in Figure 1). We analyzed feature importance by correlating individual acoustic features with embedding dimensions, performed HDBSCAN clustering [2] on the embeddings, and visualized feature distributions for concatenated and unconcatenated samples.

3. Results

3.1 Feature-Embedding Correlation

Correlation analysis between acoustic features and embedding dimensions shows that spectral and modulation-based features - particularly RASTA-filtered Linear Predictive Coding (LPC) and spectral energy coefficients - strongly influenced the spatial structure of both UMAP and t-SNE projections ($|r| > 0.7$ for top features, as shown in Table 1).

Feature (UMAP)	Corr
audspecRasta_lengthL1norm_sma_lpc0	0.726
audspecRasta_lengthL1norm_sma_lpc1	0.720
audSpec_Rfilt_sma[20]_lpc1	0.711

Table 1: *Top features by mean absolute embedding correlation.*

3.2 Clustering Analysis

The UMAP projection of acoustic features indicates a strong apparent separation between most concatenated and fluent utterances. HDBSCAN clustering in UMAP space identified 4 clusters. Clusters 0 and 1 consist almost entirely of fluent speech, while cluster 3 is $> 99\%$ concatenated and exhibits high z-scores (relative to the dataset mean) in spectral variance and roll-off features. Cluster 2, which contains both fluent and concatenated samples, shows divergent trends in RASTA-filtered LPC-based features (e.g., $+0.92$ for `lpc1` vs -0.91 for `lpc0`), indicating deviations in the spectral envelope.

This work was conducted in collaboration with Electronic Arts SEED and funded via the Sound Interactions in the Metaverse Centre for Doctoral Training and the EPSRC.

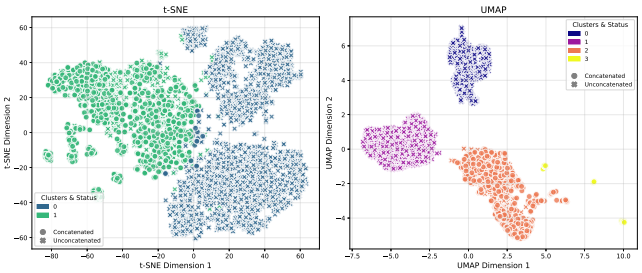

Figure 1: *t-SNE and UMAP projections of OpenSMILE ComPare_2016 features.*

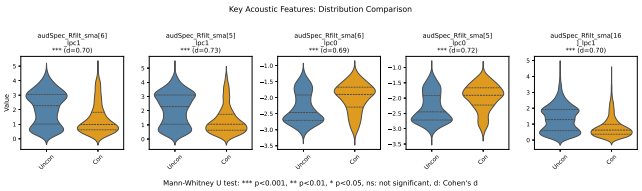

Figure 2: *Top feature distribution across concatenated and unconcatenated speech.*

3.3 Feature Distribution

We selected a subset of features with the highest correlation with UMAP dimensions and analyzed their distribution across samples and HDBSCAN clusters. As shown in Figure 2, top-ranked features display significant shifts between concatenated and fluent samples.

1. References

[1] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia (MM)*, 2010, pp. 1459–1462.

[2] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.