Table 1: Spoken Term Detection MTWV results (↑) under various distortion conditions.

Model	IV						OOV					
	-5dB	**0dB**	**5dB**	**10dB**	**15dB**	**20dB**	**-5dB**	**0dB**	**5dB**	**10dB**	**15dB**	**20dB**
Noise												
ASR Posteriors:												
HuBERT-Large	0.13	0.20	0.29	0.40	0.46	0.46	0.15	0.26	0.35	0.40	0.42	0.43
WavLM-Large	0.30	0.32	0.40	0.52	0.52	0.56	0.30	0.38	0.42	0.43	0.44	0.46
Speech Tokens:												
SpeechTokenizer	0.14	0.26	0.39	0.49	0.52	0.53	0.14	0.24	0.31	0.39	0.45	0.48
WavLM-Large	0.18	0.35	0.39	0.48	0.52	0.56	0.18	0.26	0.31	0.42	0.43	0.46
BEST-STD	0.26	0.34	0.42	0.50	0.56	0.61	0.23	0.30	0.36	0.45	0.46	0.48
Ours - Transformer	0.53	0.60	0.64	0.71	0.72	0.74	0.49	0.59	0.62	0.63	0.66	0.67
BEST-STD 2.0	**0.58**	**0.64**	**0.75**	**0.79**	**0.80**	**0.82**	**0.51**	**0.62**	**0.65**	**0.68**	**0.69**	**0.71**
Noise + Reverberation ($t_{60} = 0.7s$)												
ASR Posteriors:												
HuBERT-Large	0.02	0.07	0.11	0.21	0.23	0.25	0.07	0.11	0.17	0.23	0.25	0.27
WavLM-Large	0.10	0.19	0.21	0.33	0.35	0.36	0.16	0.21	0.28	0.32	0.36	0.38
Speech Tokens:												
SpeechTokenizer	0.05	0.11	0.16	0.17	0.21	0.21	0.06	0.12	0.16	0.21	0.21	0.23
WavLM-Large	0.07	0.14	0.21	0.25	0.36	0.39	0.09	0.15	0.21	0.22	0.26	0.27
BEST-STD	0.19	0.23	0.30	0.34	0.42	0.48	0.15	0.23	0.29	0.31	0.34	0.39
Ours - Transformer	0.40	0.51	0.54	0.59	0.60	0.62	0.41	0.47	0.53	0.56	0.58	0.60
BEST-STD 2.0	**0.46**	**0.55**	**0.62**	**0.69**	**0.70**	**0.71**	**0.41**	**0.52**	**0.57**	**0.60**	**0.62**	**0.64**

1 Additional Experimental Results

In Table 1, we provide results under various noise and reverberation conditions for the VCTK database. The VCTK corpus consists of about 44 hours of speech from 109 native and non-native English speakers, each reading approximately 400 phonetically balanced sentences. The results exhibit a trend consistent with those observed on other datasets, indicating that the proposed method consistently achieves superior accuracy compared to the baseline approaches.

2 Ablation

We analyze the impact of the two proposed components—balanced codebook constraint and data augmentation—on retrieval performance (MTWV), using in-vocabulary queries under varying noise levels. Results are shown in Table 2, evaluated under the same setup as in the main paper. All ablation experiments were carried out on the *train-clean-100* subset of the LibriSpeech dataset.

- **Model A (Baseline)**: Uses a parametric codebook trained end-to-end without augmentation or balancing. Performance drops significantly in noisy settings.
- **Model B (A + Augmentation)**: Adding data augmentation improves robustness across all SNRs, e.g., MTWV rises from 0.07 to 0.13 at -5 dB.
- **Model C (A + Balanced Codebook)**: Imposing the balanced codebook constraint yields larger gains than augmentation alone, improving clean MTWV from 0.42 to 0.74, and from 0.07 to 0.29 at -5 dB. This mitigates codebook collapse and enhances token discriminability, particularly for acoustically similar terms.
- **Model D (Balanced Codebook + Augmentation, Ours)**: Combining both components delivers the best performance across all conditions, confirming their complementary benefits and robustness under noise.

Table 2: Impact of each proposed component on MTWV in noisy conditions.

Model	**-5dB**	**0dB**	**5dB**	**10dB**	**Clean**
A: Baseline	0.07	0.09	0.13	0.21	0.42
B: A + Augmentation	0.13	0.18	0.26	0.39	0.51
C: A + Balanced Codebook	0.29	0.44	0.58	0.64	0.74
D: Balanced Codebook + Augmentation (Ours)	**0.58**	**0.64**	**0.72**	**0.75**	**0.77**

3 Qualitative Analysis

To investigate the nature of the learned tokens, we conducted a qualitative analysis by examining what each token represents. Specifically, we concatenated short audio segments (each 25ms long) that were assigned the same token. This allowed us to listen to the acoustic patterns captured by individual tokens.

Our observations reveal that these tokens are semantic in nature, consistently corresponding to syllable-like units, irrespective of speaker variation. This suggests that the learned representation captures meaningful subword structures that are more aligned with syllables than phonemes, contrary to some claims in prior work on semantic speech tokenizers.

We include example audio samples corresponding to selected tokens in the supplemental material to illustrate this behavior. These samples can be found at `https://github.com/anupsingh15/BEST-STD2.0/`