Tracking Phonetic Transitions in Self-Supervised Speech Representations

1st Kumar Kaustubh
HCLTech, Noida
kumar.kaustubh@hcltech.com

2nd Aastha Bidhen Kachhi
HCLTech, Noida
aasthabidhen.kachhi@hcltech.com

3rd Ashish Issac
HCLTech, Noida
ashishis@hcltech.com

4th Ajay Kumar Sharma
HCLTech, Noida
ajayks@hcltech.com

***Abstract*—Self-supervised learning (SSL) has enabled powerful speech representations learned directly from raw speech, yet the temporal structure encoded in these embeddings remains poorly understood. This work investigates whether phonetic segmentation signals emerge naturally in the temporal dynamics of SSL speech representations. Frame-level embeddings from pretrained SSL models (wav2vec2.0, HuBERT, and WavLM) are analyzed, and a representation-change signal is constructed by measuring cosine distance between contextual embedding averages computed over short temporal windows before and after each frame. Peaks in this signal are interpreted as candidate phonetic transitions. Using the TIMIT corpus with phoneme boundary annotations (6300 utterances), the alignment between detected peaks and phoneme boundaries is evaluated while analyzing the influence of temporal context size. Across models, representation-change peaks show strong correlations with phoneme counts in an utterance (up to 0.93) and occur close to phoneme boundaries, with median alignment errors as low as 12.7 ms. Boundary detection precision reaches 0.93 within a ±40 ms tolerance window, with F1 scores up to 0.89. These results indicate that phonetic boundary information emerges in the temporal dynamics of SSL embeddings and suggest that simple representation-transition signals provide an effective unsupervised cue for phonetic boundary detection.**

***Index Terms*—self-supervised learning, phoneme segmentation, wav2vec 2.0, HuBERT, WavLM.**

I. Introduction

Self-supervised learning (SSL) has become a central paradigm for learning speech representations directly from raw speech, enabling models to capture rich acoustic and linguistic information without requiring large labeled datasets. Recent SSL frameworks have therefore emerged as a powerful foundation for modern speech processing systems [1]. Models such as wav2vec 2.0 [2], HuBERT [3], and WavLM [4] achieve strong performance across a wide range of tasks, including automatic speech recognition, speaker recognition [5], speech emotion recognition [6], and speech synthesis [7]. Using large-scale unlabeled speech corpora, these models learn contextual representations that encode rich acoustic and linguistic information.

Speech is inherently structured in time, with linguistic units such as phonemes, syllables, and words forming a hierarchical organization [8]. Traditional speech processing systems typically segment speech into phonetic units using supervised models trained with explicit boundary annotations [9]. In contrast, SSL models are trained without phonetic supervision, yet their learned representations implicitly encode phonetic and linguistic structure [2], [3]. Understanding how such structure emerges in SSL representations may provide insights into model behavior and potentially benefit a wide range of speech technology applications.

Despite the strong performance of SSL representations across speech tasks, their temporal dynamics remain relatively underexplored. Prior studies have demonstrated that SSL embeddings encode phonetic information and support phoneme recognition or probing tasks [2], [10], [11], [12], [13], [14], [15]. However, most existing analyses focus on information contained within individual embedding frames or evaluate representations through fine-tuning on different downstream tasks [16]. Comparatively little attention has been given to how SSL embeddings evolve over time and whether transitions in these representations reflect underlying phonetic changes in the speech signal, which may be important for understanding what these models encode and learn.

This work explores the temporal behavior of SSL embeddings to examine whether phonetic transitions are reflected in their representation trajectories. If embedding dynamics capture phonetic structure, abrupt changes in the embedding space should occur near phoneme boundaries and exhibit systematic relationships with the phonetic organization of speech.

To investigate this idea, frame-level embeddings from pretrained SSL models are analyzed to identify transitions in their temporal trajectories. Using phoneme-level annotations from the TIMIT corpus [17], the relationship between detected representation transitions and phonetic structure is evaluated, and the effectiveness of these transitions as an unsupervised cue for phonetic segmentation is assessed.

The main contributions of this work are summarized as follows:

- An analysis of the temporal dynamics of self-supervised speech representations to examine whether phonetic transitions emerge in SSL embedding trajectories.
- A simple representation-change signal based on distances between contextual embedding averages for detecting candidate phonetic transitions.
- A large-scale evaluation on 6300 utterances from the TIMIT corpus using three SSL models (wav2vec 2.0, HuBERT, and WavLM), analyzing statistical relationships with phoneme structure and phoneme boundary detection performance.

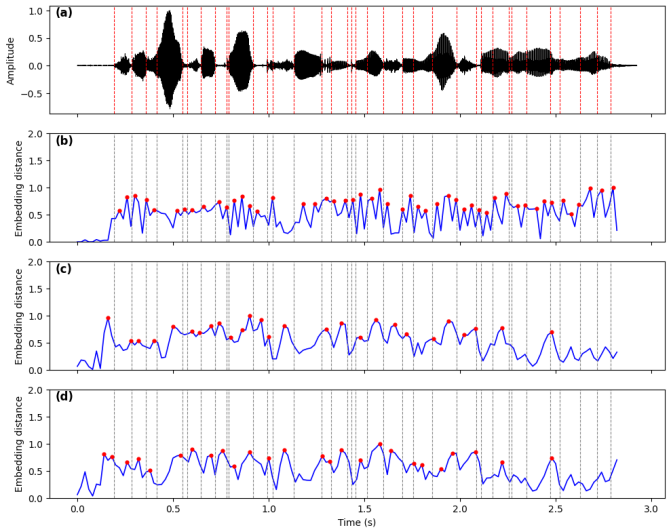

Fig. 1: (a) Waveform of the utterance *"She had your dark suit in greasy wash water all year"*. Representation-change signals computed from contextual embeddings of (b) wav2vec 2.0, (c) HuBERT, and (d) WavLM. Vertical dashed lines indicate phoneme boundaries and red markers denote detected peaks corresponding to candidate phonetic transitions.

II. Methodology

This section describes the methodology used to analyze temporal changes in self-supervised speech representations and their alignment with phoneme boundaries.

The approach consists of four stages: (1) extraction of frame-level SSL embeddings, (2) construction of a representation-change signal from embedding trajectories, (3) detection of transition peaks in this signal, and (4) evaluation of alignment between detected peaks and phoneme boundaries. These stages are described in the following subsections. An illustrative example of the representation-change signal is shown in Fig. 1 for the TIMIT utterance *SA1* spoken by speaker *FCJF0*, where peaks in the SSL embedding distance signals occur near phoneme boundaries.

Let the input speech signal be represented as a discrete-time waveform

$$x = \{x_1, x_2, \ldots, x_T\} \tag{1}$$

sampled at a rate of f_s.

The waveform is processed using pretrained SSL speech encoders that produce contextualized frame-level representations. These models typically consist of convolutional feature encoders followed by transformer-based context networks. Formally,

$$\mathbf{E} = f_\theta(x) \tag{2}$$

where $f_\theta(\cdot)$ denotes the pretrained SSL encoder with parameters θ. The resulting sequence of embedding vectors is

$$\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N\}, \tag{3}$$

where

$$\mathbf{e}_t \in \mathbb{R}^d \tag{4}$$

represents the contextualized speech representation at frame index t.

Due to the convolutional feature encoder used in SSL models, the embedding sequence is temporally subsampled relative to the input waveform. The resulting sequence $\mathbf{E}$ can therefore be interpreted as a trajectory of speech representations evolving over time in a high-dimensional embedding space.

In this study, embeddings are extracted from the final transformer layer of the pretrained SSL encoders, which provides fully contextualized speech representations. This allows the temporal dynamics of the SSL representation space to be analyzed directly. Analysis of intermediate layer representations is beyond the scope of this study.

A. Representation Change Signal

Within a phoneme segment, acoustic characteristics tend to remain relatively stable, resulting in smooth evolution of embedding vectors. In contrast, transitions between phonemes involve abrupt articulatory and acoustic changes that produce larger variations in the embedding trajectory.

To quantify these variations, representation change around each frame is measured using short temporal context windows. Let k denote the context size. The left-context representation is defined as

$$\mathbf{l}_t = \frac{1}{k} \sum_{i=1}^{k} \mathbf{e}_{t-i} \tag{5}$$

and the right-context representation as

$$\mathbf{r}_t = \frac{1}{k} \sum_{i=1}^{k} \mathbf{e}_{t+i}. \tag{6}$$

Averaging the k preceding and succeeding embeddings provides a stable estimate of the local phonetic state while excluding the center frame, enabling clearer representation of changes across potential boundaries. The representation-change signal at frame t is computed as

$$d_t = D(\mathbf{l}_t, \mathbf{r}_t) \tag{7}$$

where $D(\cdot, \cdot)$ denotes a distance function between embedding vectors. Cosine distance is primarily analyzed,

$$D_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \tag{8}$$

which measures the angular difference between embedding vectors. Cosine distance is commonly used for comparing high-dimensional embeddings and was found to produce clearer representation-change peaks than L1 or L2 distances in preliminary experiments. Computing this distance for each frame yields a one-dimensional representation-change signal whose peaks correspond to large changes in the embedding trajectory.

B. Peak Detection

Candidate phonetic boundaries are detected as local maxima in the representation-change signal. A frame index t is considered a peak if

$$d_t > d_{t-1} \quad \text{and} \quad d_t > d_{t+1}. \tag{9}$$

Detected peak indices are subsequently converted to temporal locations using the frame resolution of the SSL embeddings.

C. Boundary Alignment

Let b_j denote the ground-truth phoneme boundary timestamps obtained from the TIMIT corpus. For a detected peak occurring at time t, the temporal distance to the nearest phoneme boundary is computed as

$$\delta = \min_j |t - b_j|. \tag{10}$$

This distance is used to evaluate segmentation precision under different tolerance thresholds.

D. Utterance-Level Statistical Analysis

Beyond boundary detection, statistical relationships between representation-change peaks and phonetic structure are analyzed at the utterance level. For each utterance, the number of detected peaks N_p, the number of phonemes N_{ph}, and the utterance duration T are computed.

Peak rate and phoneme rate are then defined as

$$r_p = \frac{N_p}{T}, \qquad r_{ph} = \frac{N_{ph}}{T}. \tag{11}$$

Pearson correlation coefficients are computed across utterances to quantify relationships between representation-change peaks and phoneme counts, as well as between peak rate and phoneme rate.

III. Experimental Setup

This section describes the dataset used in the experiments, the SSL models analyzed in this study, and the evaluation protocol used to assess the relationship between representation-change peaks and phoneme boundaries.

A. Dataset

Experiments are conducted using the TIMIT corpus [17], a widely used benchmark for phonetic analysis. The corpus contains recordings from 630 speakers across multiple dialect regions of the United States, each producing ten phonetically balanced sentences, resulting in 6300 utterances. All recordings are provided at a sampling rate of f_s = 16 kHz.

TIMIT includes time-aligned phoneme annotations, which are used as ground-truth boundaries for evaluating the alignment between detected representation-change peaks and phonetic transitions.

B. SSL Models

Representations from the base variants of three SSL speech models are analyzed: wav2vec 2.0 [2], HuBERT [3], and WavLM [4]. The configurations of these models are summarized in Table I. For each utterance, frame-level embeddings are extracted from the final transformer layer of the pretrained SSL encoders.

The extracted embeddings are used to construct the representation-change signal described in Section II. Contextual embedding averages are computed using short temporal windows before and after each frame. Multiple context sizes ($k = 1, 2, 3$) are evaluated to study the influence of temporal context on segmentation performance.

Peaks in the representation-change signal are treated as candidate phonetic transitions. Peak detection applies a minimum peak height threshold of $\tau = 0.5$ and enforces a minimum separation of two frames between consecutive peaks. The threshold is selected empirically to suppress small fluctuations in the representation-change signal while preserving prominent transitions associated with phoneme boundaries; lower values increase spurious peaks while higher values miss weaker boundaries.

For the SSL models considered in this study, the convolutional feature encoder produces embeddings at a temporal resolution of 20 ms, determined by the stride of the feature encoder. Detected peak indices are converted to timestamps using this frame resolution. Alignment with phoneme boundaries is evaluated using tolerance windows of $\pm$20 ms, $\pm$30 ms, and $\pm$40 ms. Boundary detection performance is measured using precision, recall, and F1 score.

All experiments are conducted on the complete TIMIT dataset. In addition to boundary detection performance, utterance-level statistics are analyzed to examine the relationship between representation-change peaks and phonetic structure.

IV. Results

This section presents the experimental results. Statistical relationships between representation-change peaks and phoneme structure are first analyzed, followed by an evaluation of phoneme boundary detection performance.

A. Representation-Change Peak Statistics

Table II summarizes statistical relationships between detected peaks and phonetic structure across SSL models and context sizes. Across all models, strong correlations are observed between the number of detected peaks and the number of phonemes in each utterance. These correlations remain consistently high (above 0.85) for all context sizes, reaching 0.93 for WavLM. This indicates that abrupt changes in SSL embedding trajectories closely track phonetic transitions in the speech signal.

The relationship between peak rate and phoneme rate is also examined. As shown in Table II, these correlations are moderate (e.g., 0.59 for wav2vec 2.0, 0.72 for HuBERT, and

TABLE I: Configuration of the self-supervised speech models analyzed in this study.

Model	Encoder Architecture	Parameters	Pretraining Data	Hours	Embedding Dimension
wav2vec2-base	CNN + Transformer	95M	LibriSpeech	960 h	768
HuBERT-base	CNN + Transformer	95M	LibriSpeech	960 h	768
WavLM-base	CNN + Transformer	94M	Libri-Light + VoxPopuli + GigaSpeech	94k h	768

0.73 for WavLM at $k = 1$), indicating that representation-change peaks are influenced by speaking rate but are not determined solely by speech tempo and primarily reflect local phonetic transitions.

Temporal alignment between detected peaks and ground-truth phoneme boundaries is also highly precise. Median alignment errors reported in Table II range from 20 ms for wav2vec 2.0 to 12.75 ms for WavLM. These values are close to the intrinsic 20 ms temporal resolution of SSL embeddings, indicating that detected transitions provide temporally accurate indicators of phoneme boundaries. Among the evaluated models, WavLM consistently achieves the lowest alignment error, suggesting that its representations capture phonetic transitions with sharper temporal structure.

B. Unsupervised Boundary Detection Performance

Table III evaluates the alignment between detected peaks and phoneme boundaries under different tolerance windows using the context size that produces the best precision for each model ($k = 1$), along with the corresponding recall and F1 scores. Precision is emphasized because peaks in the representation-change signal are interpreted as candidate phonetic transitions. High precision therefore indicates that detected peaks closely correspond to true phoneme boundaries, while recall reflects the proportion of ground-truth boundaries successfully detected.

Across all SSL models, cosine distance between contextual embeddings yields reliable boundary detection performance. HuBERT and WavLM achieve the highest segmentation precision, reaching 0.93 within a ± 40 ms tolerance window, while wav2vec 2.0 achieves a precision of 0.80 under the same condition. Corresponding F1 scores reach up to 0.89.

The influence of context size and model architecture on boundary detection precision is illustrated in Fig. 2. Increasing the context size averages neighboring embeddings and introduces temporal smoothing, which can reduce sensitivity to sharp representation changes associated with phoneme transitions. HuBERT and WavLM consistently achieve higher precision than wav2vec 2.0 across tolerance windows, indicating that their learned representations capture phonetic transitions with clearer temporal boundaries.

V. Discussion

The results suggest that phonetic transitions are reflected in the temporal dynamics of SSL embedding trajectories. Within a phoneme segment, acoustic properties tend to remain relatively stable, leading to smooth evolution of contextualized embeddings. In contrast, transitions between phonemes involve changes in articulatory configuration that produce distinct acoustic patterns. These acoustic changes are captured by the SSL models as movements in the embedding space, resulting in representation-change peaks that tend to occur near phoneme boundaries.

TABLE II: Representation-change peak statistics across SSL models for different context sizes k.

Model	k	Count Corr.	Rate Corr.	Median Error (ms)
wav2vec 2.0	1	0.91	0.59	20.00
	2	0.85	0.48	19.19
	3	0.76	0.43	18.00
HuBERT	1	0.91	0.72	14.06
	2	0.93	0.69	14.38
	3	0.91	0.57	14.38
WavLM	1	0.93	0.73	12.75
	2	0.93	0.67	13.69
	3	0.91	0.58	13.94

TABLE III: Boundary detection performance for representation-change peaks using cosine distance. Results use the best context size per model and are evaluated under different tolerance windows.

Model	k	Tolerance (ms)	Precision	Recall	F1
wav2vec 2.0	1	± 20	0.50	0.61	0.55
		± 30	0.68	0.84	0.75
		± 40	0.80	0.99	0.89
HuBERT	1	± 20	0.67	0.47	0.55
		± 30	0.85	0.60	0.71
		± 40	0.93	0.66	0.77
WavLM	1	± 20	0.68	0.53	0.59
		± 30	0.84	0.65	0.73
		± 40	0.93	0.71	0.80

This behavior may also be related to the training objectives of SSL speech models. Masked prediction encourages the model to infer missing speech segments from surrounding context, promoting temporally consistent representations within phoneme regions while amplifying changes across phonetic transitions. Models such as HuBERT and WavLM further incorporate discrete unit prediction during pretraining, which encourages the representation space to organize around phonetic-like acoustic categories while WavLM's additional denoising objective and larger pretraining data further sharpen its phonetic transitions. This training objective may contribute to sharper transitions in their embedding trajectories and clearer boundary-related changes in the representation-change signal.

The influence of context size further highlights the largely local nature of phonetic information in SSL embeddings. Small context windows emphasize short-term representation changes and preserve abrupt transitions associated with phoneme boundaries, whereas larger contexts average neighboring embeddings and smooth local variations. This behavior suggests that phonetic transitions appear as relatively sharp movements in SSL representation space rather than gradual shifts over longer temporal spans.

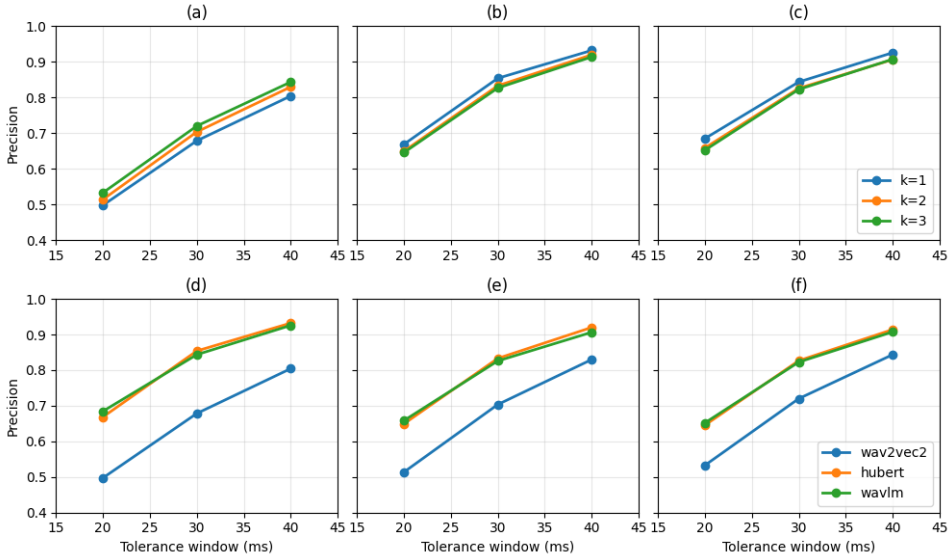

Fig. 2: Precision versus tolerance window for boundary detection using cosine distance between contextual embedding averages. Effect of context size k for (a) wav2vec 2.0, (b) HuBERT, and (c) WavLM and comparison of SSL models for (d) $k = 1$, (e) $k = 2$, and (f) $k = 3$.

VI. Conclusion

This work investigated whether phonetic segmentation signals emerge in the temporal dynamics of self-supervised speech representations. Frame-level embeddings from wav2vec 2.0, HuBERT, and WavLM were analyzed to construct a representation-change signal that captures transitions in SSL embedding trajectories. Experiments on the TIMIT corpus demonstrate strong alignment between detected peaks and phoneme boundaries, with peak–phoneme count correlations up to 0.93 and boundary precision reaching 0.93 within a ± 40 ms tolerance window.

The results further reveal differences in temporal behavior across SSL models and indicate that phonetic boundary cues are primarily reflected in local changes in embedding trajectories. These findings suggest that the temporal dynamics of SSL embeddings provide an interpretable signal for analyzing phonetic structure in speech. Future work may extend this analysis to Indic and other low-resource languages, as well as to more diverse linguistic settings including spontaneous and conversational speech, and explore whether similar representation-transition signals, including those from intermediate layers which are known to capture richer phonetic information, can support unsupervised segmentation at higher linguistic levels such as syllables or words.

References

[1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[5] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7967–7971.

[6] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," 2021. [Online]. Available: https://arxiv.org/abs/2104.03502

[7] S. Wang, G. E. Henter, J. Gustafson, and E. Szekely, "On the use of self-supervised speech representations in spontaneous speech synthesis," *arXiv preprint arXiv:2307.05132*, 2023.

[8] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[9] S. Young, G. Evermann *et al.*, *The HTK Book*. Cambridge University Engineering Department, 2002.

[10] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *ICASSP*, 2021.

[11] M. Riviere *et al.*, "Unsupervised pretraining transfers well across languages," in *ICASSP*, 2020.

[12] H. Ji, T. Patel, and O. Scharenborg, "Predicting within and across language phoneme recognition performance of self-supervised learning speech pre-trained models," *arXiv preprint arXiv:2206.12489*, 2022.

[13] F. Kreuk, J. Keshet, and Y. Adi, "Self-supervised contrastive learning for unsupervised phoneme segmentation," in *Proc. Interspeech*, 2020, pp. 3427–3431.

[14] L. Strgar and D. Harwath, "Phoneme segmentation using self-supervised speech models," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar, 2023, pp. 1–8.

[15] A. Pasad, C.-M. Chien, S. Settle, and K. Livescu, "What do self-supervised speech models know about words?" *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 372–391, 2024.

[16] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Timit acoustic-phonetic continuous speech corpus," 1993.