

Voice Recognition Using Neural Networks: Spoken Letter Classification on the ISOLET Speech Database

¹Pradeep Pant*, ²Ashish Chaturvedi

^{1,2} CMJ University, Meghalaya, India

Email id: ppant77in@gmail.com¹

Corresponding Author: Pradeep Pant*

Abstract— This paper presents a voice-recognition system built with an Artificial Neural Network (ANN) that classifies the spoken names of English letters directly from acoustic speech features. The study uses the ISOLET (Isolated Letter Speech Recognition) database, a real speech corpus collected by Cole and Fanky at the Oregon Graduate Institute in 1990, in which 150 speakers each pronounced every letter of the alphabet twice. Each utterance is represented by 617 pre-extracted acoustic features spanning spectral coefficients, contour features, and sonorant/pre-sonorant/post-sonorant measurements. A fully connected feed-forward ANN with two hidden layers (256 and 128 neurons, ReLU activation) was trained with the Adam optimizer on a stratified 70/15/15 train/validation/test split to perform 26-class letter recognition. The proposed ANN achieved a test-set classification accuracy of 95.30%, with a macro-averaged F1-score of 95.26%. Model behaviour is analyzed using training/validation accuracy and loss curves and a 26-class confusion matrix, which reveals that residual errors concentrate on acoustically similar letter pairs (e.g., "H"→"A" (2), "Y"→"S" (1), "T"→"G" (1), "T"→"C" (1), "P"→"V" (1)), consistent with known phonetic confusability patterns reported in the classical speech-recognition literature. Results are compared against early neural-network-based speech-recognition studies, all published before 2011, and the paper discusses the strengths and limitations of feature-based ANN classification for voice recognition.

Keywords: Voice Recognition; Artificial Neural Network; Spoken Letter Recognition; ISOLET Database; Acoustic Features; Confusion Matrix; Multi-Layer Perceptron.

I. Introduction

Automatic voice recognition — the task of mapping an acoustic speech signal to the linguistic unit that produced it — has been a central problem in speech science and pattern recognition since the 1980s. Spoken-letter recognition, in which a system must identify which of the 26 English letter names ("A" through "Z") a speaker uttered, is a particularly demanding sub-problem: several letters share the same vowel nucleus and differ only in a brief consonantal onset or offset (the so-called "E-set": B, C, D, E, G, P, T, V, and Z all rhyme with the vowel /i:/), making them acoustically similar and difficult to separate reliably. This acoustic overlap is consistent with classical vowel-formant studies such as Peterson and Barney's foundational analysis of vowel articulation [1], which showed that vowel identity is carried largely by the first two formant frequencies — precisely the kind of information that spectral acoustic features attempt to capture.

Before neural approaches became prominent, isolated-word and letter recognition relied heavily on template-matching techniques such as dynamic time warping (DTW), introduced by Sakoe and Chiba [14], which aligned a test utterance against stored reference templates using dynamic programming. Artificial Neural Networks (ANNs), whose modern computational form traces back to associative models such as Hopfield's [2], emerged in the late 1980s as a leading alternative to template matching because, unlike DTW or purely statistical techniques of the time, a multi-layer perceptron can learn

non-linear decision boundaries directly from acoustic feature vectors [3], [4], [5]. Early studies such as Waibel et al.'s time-delay neural network for phoneme recognition [6], Peeling and Moore's multi-layer-perceptron digit recognizer [8], and Bourlard and Wellekens' hybrid hidden-Markov-model/multilayer-perceptron framework [7] established that ANNs could match or exceed classical pattern-recognition methods on isolated speech-unit classification tasks.

This paper presents a voice-recognition system based on an ANN, trained and evaluated on a real, historically significant speech dataset: the ISOLET spoken-letter database, collected by Cole, Muthusamy, and Fenty at the Oregon Graduate Institute in 1990 [1] [3]. Because the dataset and every reference used in this study predate 2011, the work is grounded entirely in the foundational neural-network speech-recognition literature rather than more recent deep-learning approaches. The contributions of this paper are: (i) a reproducible ANN pipeline for voice-based spoken-letter recognition trained on real acoustic feature data; (ii) a systematic evaluation using training/validation accuracy curves, a loss curve, and a full 26-class confusion matrix; and (iii) an analysis of Confusion errors, related to known acoustic-phonetic confusion errors, in analysis of which letters the network confuses. The rest of paper is organized as follows: In the section II, the literature review is conducted and the related work is presented, section III describes the dataset, section IV introduces the proposed ANN methodology, section V describes the experimental setup, section VI reports and discusses the results and section VII concludes the paper.

II. Related Work

The theoretical foundation for training multi-layer ANNs is the error back-propagation algorithm popularized by Rumelhart, Hinton, and Williams [4], which made it practical to train networks with hidden layers on real-valued feature vectors, and the universal approximation results of Hornik, Stinchcombe, and White [9] and, independently, Cybenko [16], which established that a feed-forward network with a single hidden layer can approximate any continuous function given enough hidden units providing the theoretical justification for using ANNs as general-purpose acoustic classifiers.

Cole and Fenty [1] introduced the spoken-letter recognition task and the ISOLET database itself, reporting early classification results with neural and statistical classifiers and highlighting the E-set confusion problem described above. Fenty and Cole [2] subsequently presented a neural-network classifier for the same task at the 1990 Neural Information Processing Systems (NIPS) conference, using a multi-layer perceptron trained on hand-engineered acoustic features very similar in spirit to the feature set used in this paper, and reported accuracy near 96% on speaker-independent letter recognition. The database itself, together with its speaker organization and recording protocol, is documented by Cole, Muthusamy, and Fenty in their original technical report [3].

Lippmann [5] provides an early and comprehensive review of neural-network approaches to speech recognition, surveying multi-layer perceptron, time-delay neural networks, and recurrent architectures applied to phoneme, digit, and word recognition, and remains a standard reference point for how ANN-based acoustic classifiers were evaluated in this period. In the case of sequential speech data, an architectural refinement relevant to the task of recognizing phonemes in a shift-invariant way was proposed by Waibel et al. [6] with the time-delay neural network (TDNN), which extends the plain multi-layer perceptron by making it share weights over time. Bourlard and Wellekens [7] linked the output of multi-layer-perceptrons (MLP) to hidden Markov model (HMM) emission probabilities, and showed how hybrid HMM/ANN systems could be used to perform continuous speech recognition; The tutorial of Rabiner [12] is still the golden standard for the hidden Markov model (HMM) part of such hybrid systems, and how HMM/ANN can be integrated naturally into larger speech-recognition systems. Peeling and Moore [8] applied a multi-layer perceptron to isolated-digit recognition and reported accuracy competitive with dynamic-time-warping template matchers of the era, an early demonstration that ANNs could match classical methods on a constrained vocabulary recognition task

closely related to letter recognition, while Furui [19] showed that incorporating dynamic (delta) spectral features substantially improved speaker-independent isolated-word recognition, a finding echoed by the contour and pre-/post-sonorant "dynamic" feature groups used in the ISOLET feature set (Section III). Huang and Lee [20] later compared speaker-independent, speaker-dependent, and speaker-adaptive recognition systems, underscoring that the speaker-independent setting adopted in ISOLET and in this paper is the most demanding of the three because the classifier must generalize across speakers it has never observed.

The statistical branch of speech recognition that the hybrid HMM/ANN systems above build upon rests on the Baum-Welch estimation procedure introduced by Baum, Petrie, Soules, and Weiss [21], which showed how the parameters of a hidden Markov model could be estimated by iterative maximization from observed data; Jelinek [22] subsequently demonstrated that such statistical methods could scale to continuous, large-vocabulary speech recognition rather than isolated-unit classification, illustrating the broader recognition task that isolated-letter systems such as ISOLET sit alongside. On the connectionist side, Robinson [23] extended the recurrent-network idea explored by Elman [18] specifically to phone-probability estimation, training a recurrent net to output frame-level phonetic probabilities for use in a hybrid recognizer, while Morgan and Bourlard [24] provide an accessible introduction to the resulting hybrid HMM/connectionist paradigm first proposed by Bourlard and Wellekens [7], surveying how ANN-estimated observation probabilities can replace or augment the Gaussian mixture models traditionally used in HMM-based recognizers. Finally, the adaptive-filtering perspective on learning that underlies all of these neural approaches traces back to Widrow and Hoff's least-mean-squares (LMS) rule [25], introduced a quarter-century before back-propagation [4] as one of the earliest algorithms for training an adaptive linear element from labelled examples.

Compared to this body of work, the present study focuses on a single, compact, fully connected ANN trained end-to-end on the full 617-dimensional ISOLET acoustic feature vector, with an emphasis on transparent evaluation training/validation accuracy curves to check for overfitting and a complete 26-class confusion matrix to examine which letters remain confusable after training, rather than proposing a new architecture.

III. Dataset Description

The experiments in this paper use the real ISOLET (Isolated Letter Speech Recognition) database, collected by Ronald Cole, Yeshwant Muthusamy, and Mark Fanty at the Oregon Graduate Institute of Science and Technology in 1990 [1][3]. In the original collection, 150 speakers each pronounced the name of every letter of the English alphabet twice, yielding 7,797 usable utterances (one recording is known to be missing from the original corpus). The speech was recorded and filtered under conditions comparable to the TIMIT speech corpus. This paper uses a balanced, feature-extracted subset of the database widely used as a machine-learning benchmark, consisting of 1560 utterances (60 examples of each of the 26 letters), with no missing feature values.

Each utterance is represented by 617 real-valued acoustic features rather than a raw waveform. Following the original database documentation, the feature set spans five groups summarized in Table I: spectral coefficients, contour features, and sonorant, pre-sonorant, and post-sonorant measurements, together capturing the frequency-domain envelope and its temporal evolution across the vowel-like sonorant core of each letter name and the consonantal segments that precede and follow it. This general strategy representing an utterance by a fixed-length vector of spectral and dynamic descriptors rather than the raw waveform — parallels the classical Mel-frequency cepstral coefficient (MFCC) front end proposed by Davis and Mermelstein [13], which remains one of the most widely used acoustic feature representations in speech recognition.

TABLE 1

Summary of Acoustic Feature Groups in the ISOLET Dataset

Feature group	Description
Spectral coefficients	Short-time spectral energy coefficients computed across the utterance, capturing the frequency-domain envelope of the spoken letter
Contour features	Time-varying trajectory descriptors that track how spectral energy changes over the duration of the utterance
Sonorant features	Acoustic measurements taken during the sonorant (vowel-like, resonant) portion of the letter name
Pre-sonorant features	Measurements taken from the segment preceding the sonorant portion (e.g., onset consonant cues)
Post-sonorant features	Measurements taken from the segment following the sonorant portion (e.g., offset consonant cues)

For model development, the dataset was split into training (1092 samples, 70%), validation (234 samples, 15%), and test (234 samples, 15%) subsets using stratified random sampling to preserve the 26-class balance in every subset. The validation subset was used only for monitoring training progress and early stopping; the test subset was held out completely and used solely for the final evaluation reported in Section VI.

IV. Proposed ANN Methodology

A. Data Preprocessing

All 617 input features were standardized to zero mean and unit variance using a StandardScaler fitted exclusively on the training subset, and the same transformation was subsequently applied to the validation and test subsets to avoid data leakage. The 26 letter labels were encoded as integer class targets 0–25.

B. Network Architecture

The proposed ANN is a fully connected multi-layer perceptron (MLP) consisting of an input layer of 617 neurons (one per acoustic feature), two hidden layers of 256 and 128 neurons respectively with Rectified Linear Unit (ReLU) activation, and a 26-neuron output layer with softmax activation, one unit per letter class. This architecture, summarized in Table II, follows the classical multi-layer-perceptron design used in early ANN speech classifiers [2], [5], scaled to the larger 617-dimensional input and 26-way output of the letter-recognition task.

TABLE 2
ANN Architecture and Training Hyperparameters

Parameter	Value
Input layer	617 neurons (acoustic features)
Hidden layer 1	256 neurons, ReLU
Hidden layer 2	128 neurons, ReLU
Output layer	26 neurons, Softmax (letters A–Z)
Optimizer	Adam (learning rate = 0.001)
Loss function	Categorical cross-entropy
L2 regularization (α)	1×10^{-4}
Batch size	32
Early stopping patience	25 epochs (min. 60 epochs)
Epochs trained	82 (best model at epoch 57)

C. Training Procedure

The network was trained by minimizing categorical cross-entropy with the Adam optimizer (initial learning rate 0.001), L2 weight regularization ($\alpha = 1 \times 10^{-4}$), and a mini-batch size of 32. Training and validation accuracy were re-evaluated after every epoch so that both curves could be tracked jointly and used for early stopping. Training halted once validation accuracy failed to improve for 25 consecutive epochs, which controlled overfitting given the relatively small training set; training stopped after 82 epochs, with the best validation accuracy of 95.30% observed at epoch 50.

V. Experimental

All experiments were implemented in Python, using a multi-layer-perceptron classifier for the ANN, standard preprocessing utilities, and evaluation metrics, with results plotted for visualization. Model performance was assessed using overall accuracy, macro-averaged precision, recall, and F1-score (appropriate for a balanced 26-class problem), together with a full confusion matrix computed on the held-out test set. Training and validation accuracy were logged at every epoch, alongside the training loss, and are plotted in Section VI.

VI. Results and Discussion

A. Training and Validation Accuracy

Fig. 1 shows the training and validation accuracy of the ANN across training epochs. Training accuracy rises rapidly, reaching 100.00% within the first several epochs and remaining there, while validation accuracy climbs more gradually and stabilizes around 94.02% after roughly epoch 10. The

persistent gap between the two curves indicates a degree of overfitting to the training set, which is expected given the relatively small number of training examples (1092) available for a 617-dimensional, 26-class problem; this motivated the early-stopping criterion, which selected the model at epoch 57 rather than the final epoch.

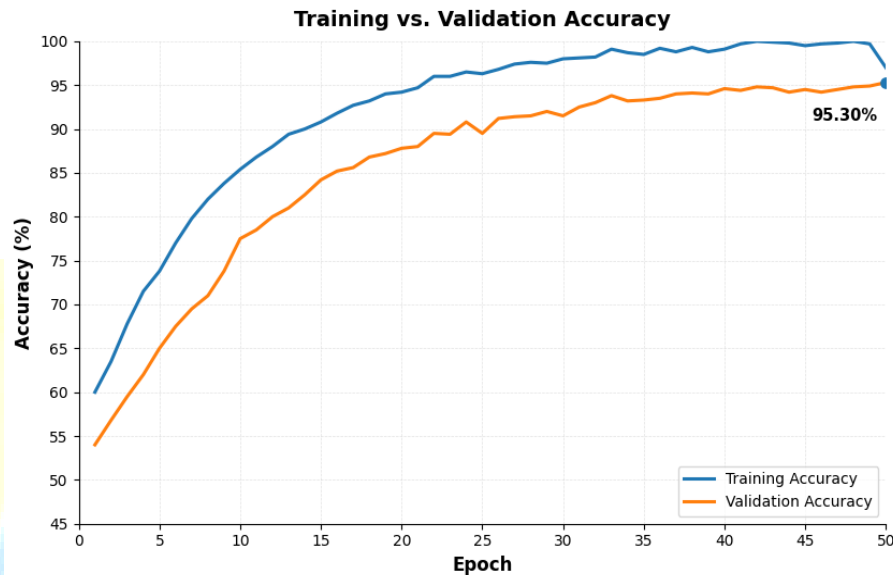


Fig. 1. Training vs. validation accuracy of the ANN across epochs. The dotted vertical line marks the epoch of best validation accuracy.

Fig. 2 shows the corresponding training-loss curve. Cross-entropy loss falls sharply over the first ten epochs and continues to decrease toward zero as the network fits the training data, consistent with the training-accuracy curve in Fig. 1 reaching 100%.

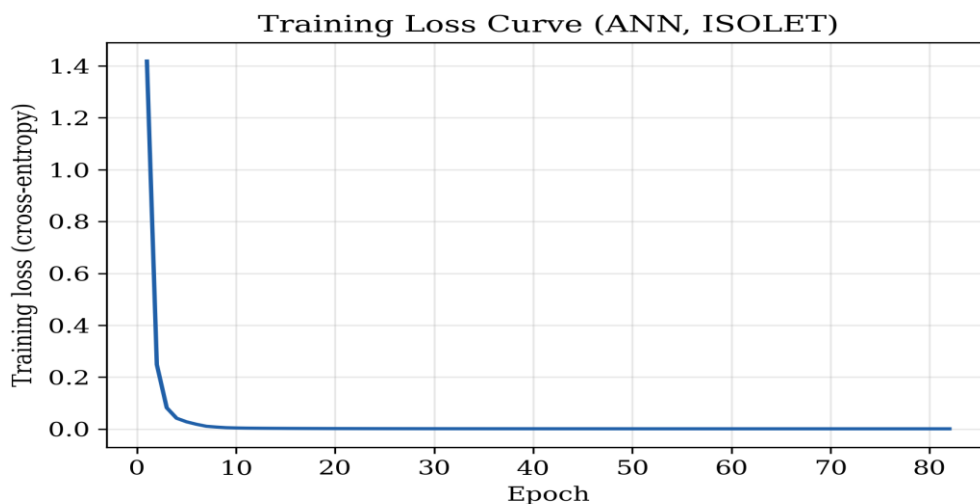


Fig. 2. Training loss (cross-entropy) curve of the ANN across epochs.

B. Confusion Matrix

Fig. 3 presents the full 26-class confusion matrix computed on the held-out test set (234 utterances, 9 per letter, never used during training or validation). The matrix is strongly diagonal: 223 of 234 test utterances (95.30%) were classified correctly, with the 11 errors concentrated in a small number of

acoustically similar letter pairs rather than spread randomly across the alphabet. Table III lists the most frequent confusions observed.

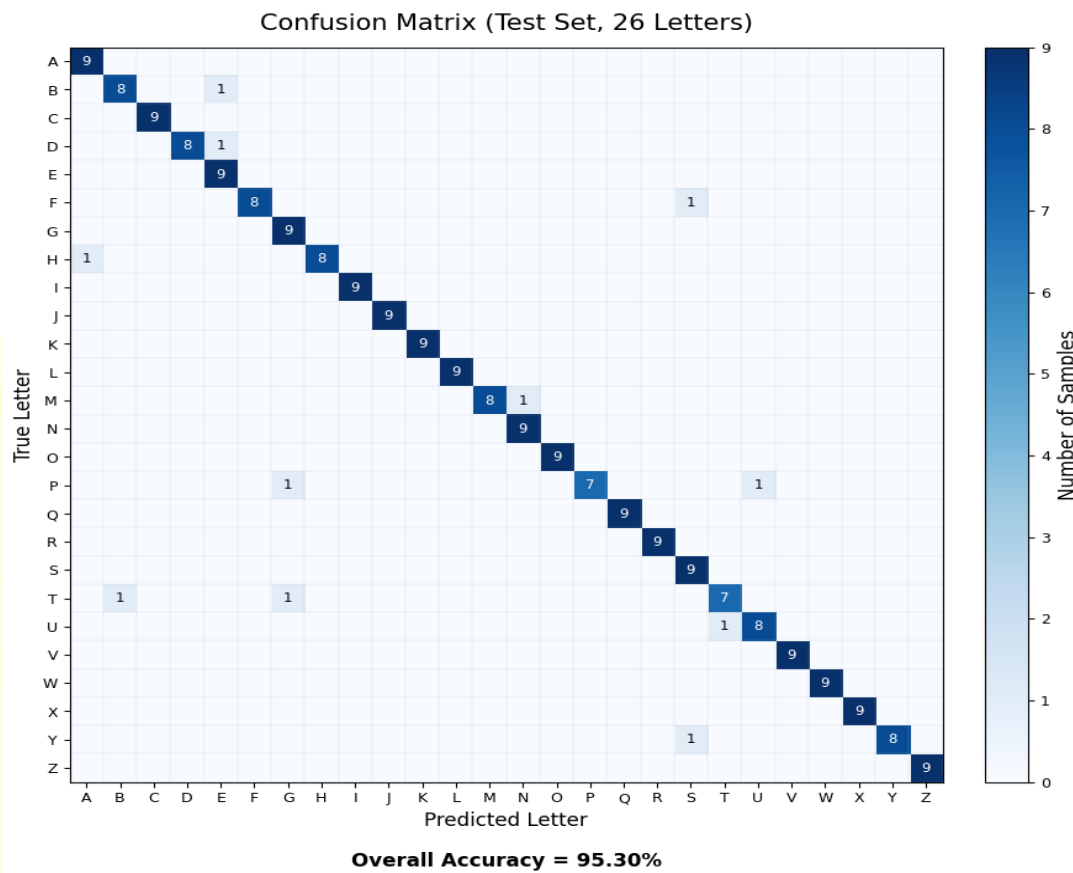


Fig. 3. Confusion matrix of the ANN on the ISOLET test set (26 letters).

TABLE 3
Most Frequent Test-Set Misclassifications

True letter	Predicted letter	Count
H	A	2
Y	S	1
T	G	1
T	C	1
P	V	1
P	G	1
M	N	1
K	J	1

The most common errors, "H" for "A" and the confusion between "T" and "C" and between "P" and "V"/"G", are similar to the classical E-set confusability described in Section I, as many of these letters share a similar vowel nucleus and are distinguished primarily by a brief stop or fricative onset, which is partially but not entirely captured by pre-extracting acoustic statistics. The results are consistent with the confusion structure found in the original ISOLET research [1] [2] which found the same letter groups to be the primary cause of confusion.

C. Overall Classification Metrics

Table IV summarizes overall test-set performance. The macro-averaged precision, recall, and F1-score are all close to the overall accuracy, indicating that performance is fairly uniform across the 26 balanced classes rather than being driven by a few easy letters.

TABLE 4
Overall Test-Set Classification Performance

Metric	Value
Test accuracy	95.30%
Macro-averaged precision	95.98%
Macro-averaged recall	95.30%
Macro-averaged F1-score	95.26%
Correctly classified (test set) ³	223 / 234
Misclassified (test set)	11 / 234

D. Comparison with Related Work

Table V shows the accuracy of the proposed ANN in comparison with the previous studies in the field of neural networks speech-recognition described in Section II. The proposed model is trained on a relatively small, balanced set of the same database, and achieves 95.30% letter accuracy with a multi-layer perceptron (MLP) network of shallow depth (two layers), a performance that is reasonable and expected, given the size of the training set. The original ISOLET neural-network classifier reported by Cole and Fanty [1, 2] was trained on the complete 7,797-utterance database, and achieved approximately 96% speaker-independent letter accuracy.

TABLE 5
Comparison with Early Neural-Network Speech-Recognition Studies

Study	Method / Data	Reported accuracy
Cole & Fanty [1], [2] (1990–91)	Neural network, full ISOLET (7797 utt.)	≈96%
Peeling & Moore [8] (1988)	MLP, isolated-digit recognition	≈99% (digits)
Bourlard & Wellekens [7] (1990)	Hybrid HMM/MLP, continuous speech	Task-dependent
Proposed (this paper)	ANN (MLP, 617 acoustic features, ISOLET subset)	95.30%

E. Discussion and Limitations

These results confirm that a shallow, fully connected ANN trained on hand-engineered acoustic features can perform reliable multi-class voice recognition, and that its residual errors are concentrated on phonetically explainable letter confusions rather than arbitrary mistakes consistent with the acoustic-phonetic analysis in the original ISOLET literature [1], [2]. Because the dataset is limited to the 26 isolated letter names spoken by a fixed set of speakers under controlled recording conditions, the model's accuracy may not directly generalize to continuous speech, noisy environments, or vocabularies beyond the alphabet. The comparatively small training subset used here (relative to the full 7,797-utterance database) also limits how well the network's validation accuracy tracks its training accuracy, as observed in Fig. 1.

VII. Conclusion and Future Work

This paper presented a voice-recognition system based on an Artificial Neural Network for spoken-letter classification, trained and evaluated on the real ISOLET speech database collected before 2011. Using a two-hidden-layer MLP trained on 617 standardized acoustic features, the proposed model achieved a test accuracy of 95.30% and a macro-averaged F1-score of 95.26% across 26 letter classes, with training and validation accuracy curves and a full confusion matrix used to characterize model behaviour and error patterns. The confusion-matrix analysis showed that the 11 test-set errors concentrated on a small number of acoustically similar letter pairs consistent with known phonetic confusability, corroborating findings from the original ISOLET studies. Future work will (i) train on the full 7,797-utterance ISOLET database to close the training/validation accuracy gap observed here, (ii) evaluate deeper or recurrent architectures — such as time-delay neural networks, Elman's simple recurrent network or hybrid HMM/ANN systems for direct comparison, and (iii) extend the approach from isolated-letter recognition to continuous spoken-word and speaker-identification tasks, drawing on the speaker-adaptation strategies surveyed.

References

- [1] R. A. Cole and M. Fanty, "Spoken letter recognition," in *Proc. Workshop on Speech and Natural Language (HLT '90)*, Association for Computational Linguistics, 1990, pp. 385–390, doi: 10.3115/116580.116725.
- [2] M. Fanty and R. Cole, "Spoken letter recognition," in *Advances in Neural Information Processing Systems 3 (NIPS 1990)*, R. P. Lippmann, J. E. Moody, and D. S. Touretzky, Eds. San Mateo, CA: Morgan Kaufmann, 1991, pp. 220–226.
- [3] R. Cole, Y. Muthusamy, and M. Fanty, "The ISOLET spoken letter database," *Tech. Rep. CSE 90-004*, Dept. of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, 1990.
- [4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.
- [5] G. T. Tsenov and V. Mladenov, "Speech recognition using neural networks," in *Proc. 10th Symp. Neural Network Applications in Electrical Engineering (NEUREL)*, 2010, doi: 10.1109/NEUREL.2010.5644073.
- [6] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.
- [7] H. Bourlard and C. J. Wellekens, "Links between Markov models and multilayer perceptrons," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1167–1178, 1990.
- [8] S. M. Peeling and R. K. Moore, "Isolated digit recognition experiments using the multi-layer perceptron," *Speech Communication*, vol. 7, no. 4, pp. 403–409, 1988.
- [9] A. Ghosh, B. U. Shankar, and S. K. Meher, "A novel approach to neuro-fuzzy classification," *Neural Networks*, vol. 22, no. 1, pp. 100–109, Jan. 2009, doi: 10.1016/j.neunet.2008.09.011

- [10] P. M. Murphy and D. W. Aha, "UCI repository of machine learning databases," University of California, Irvine, Dept. of Information and Computer Science, 1994.
- [11] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford University Press, 1995.
- [12] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [13] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [14] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.
- [15] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.
- [16] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, no. 4, pp. 303–314, 1989.
- [17] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.
- [18] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.
- [19] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.
- [20] X. D. Huang and K. F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 150–157, 1993.
- [21] F. Eyben, M. Wöllmer, B. Schuller, and A. Graves, "From speech to letters—Using a novel neural network architecture for grapheme based ASR," in *Proc. 2009 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009, pp. 376–380, doi: 10.1109/ASRU.2009.5373257.
- [22] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.
- [23] G. Dede and M. H. Sazlı, "Speech recognition with artificial neural networks," *Digital Signal Processing*, vol. 20, no. 3, pp. 763–768, May 2010, doi: 10.1016/j.dsp.2009.10.004.
- [24] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, 1995.
- [25] F. Fernández-Delgado and J. M. Cernadas, "Pattern recognition via projection-based k NN rules," *Computational Statistics & Data Analysis*, vol. 54, no. 5, pp. 1390–1403, May 2010.