Speech-to-Visualization: Toward End-to-End Speech-Driven Data Visualization Generation from Natural Language Questions

Haodi Zhang[1], Xinhe Zhang[1], Jihua Zhou[1], Kaishun Wu[4], Yuanfeng Song[2](✉), and Raymond Chi-Wing Wong[3]

[1] Shenzhen University, Shenzhen, China
[2] AI Group, WeBank Co., Ltd, Shenzhen, China
[3] The Hong Kong University of Science and Technology, Hong Kong, China
[4] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

Abstract. Data visualization (DV) has evolved rapidly, transforming intricate datasets into accessible visual representations. However, the intricate grammar of DV languages, such as Vega-Lite, presents a substantial barrier for beginners and users without technical backgrounds. To address this challenge, extensive research has focused on developing models that can translate natural language questions (NLQs) into DV languages, a process formally known as text-to-visualization in the field.

With the recent development of speech-related technologies, particularly Acoustic Speech Recognition (ASR), voice-based interaction has become a growing trend in real-world applications. In this paper, we introduce speech-to-vis, a novel task that translates speech-form NLQs into data visualizations. To address the scarcity of relevant datasets, we present *SpeechNVBench*, the first manually annotated dataset specifically designed for this field. Our research reveals that the intuitive cascaded approach (i.e., ASR followed by text-to-vis) suffers from error propagation issues, where small errors in earlier stages lead to larger errors in subsequent stages. In response, we introduce *SpeechVisNet*, the first end-to-end neural architecture that directly translates speech-form NLQs into DVs. SpeechVisNet incorporates advanced structures like a DV-aware decoder to ensure reliable output. Furthermore, to mitigate the modality gap between speech-modality questions and text-modality data schema, we explore bridging techniques to align them. Experimentation on our proposed dataset demonstrates SpeechVisNet's competitive edge against various strong baselines. This work aims to drive innovation in human-machine interfaces, enhancing the efficiency and accessibility of DV tools across various domains.

Keywords: Data Visualization· Data Analysis· Speech-Driven Visualization System· Neural Architecture· Speech-to-Visualization

1 Introduction

With the rapid growth of available data in today's digital world, the capacity of transforming complex data into meaningful visual representations has become essential for rational decision-making and eff communication. Data visualization (DV) is a cornerstone in

this process, leveraging visual elements such as bar charts, scatter plots, and histograms to convey data information to readers in a straight-forward and intuitive manner [24]. By enhancing the comprehension of data, DV enables users to grasp intricate concepts and patterns with greater ease and clarity. Recognizing its transformative impact, DV plays an indispensable role in a wide range of academic and commercial fields, including but not limited to data mining, databases, recommendation systems, and data analysis [17, 23, 34, 41]. In recent years, there have been numerous DV-related studies published in top database conferences and journals such as ICDE [28, 15] and SIGMOD [33, 18], VLDB [35, 39], and TKDE [42, 17]. For example, Sevi [32], published in SIGMOD'22, proposed an automatic DV generation system that processes natural language questions in speech form.

The creation of DVs is usually achieved by composing specifications using the DV languages. These DV languages, such as Vega-Lite [26], ggplot2 [36], ZQL [27], ECharts [16], and VizQL [13], defined in form of complicated grammars, where a json configuration file can be executed to produce a visualization chart. They empower professionals and seasoned scholars to craft sophisticated visualizations tailored to their needs, while also presenting substantial challenges for common users due to the complexity and steep learning curves. To bridge the gap, text-to-visualization (*text-to-vis*) techniques have been proposed to automate the translation of natural language questions (NLQs) into DVs, unlocking the power of databases and visualization systems for users with limited technical skills [18]. In this context, a significant body of work has contributed to the field's progress [18, 31].

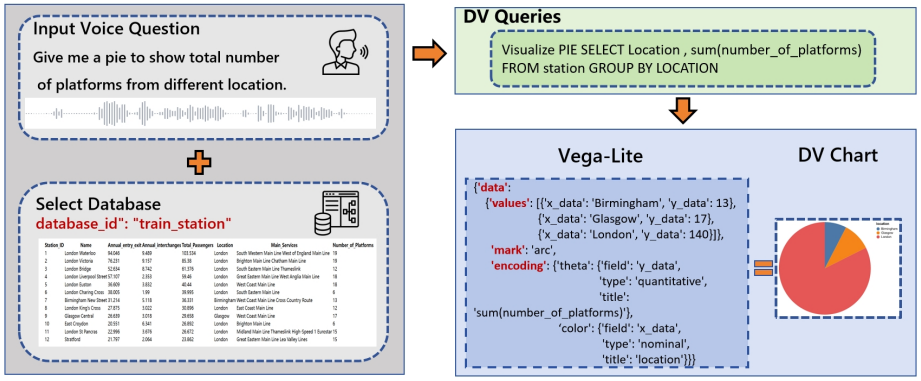


Fig. 1: Speech-to-Vis Process: From Voice Query to Data Visualization. This chart illustrates the workflow of converting a speech-based natural language question and database schema into a corresponding data visualization. It demonstrates how spoken queries can be transformed into visual representations of data, bridging the gap between verbal communication and visual analytics.

On the other hand, with the widespread utilization of smartphones and tablets, applications like voice search, AI assistants, and chatbots have gained popularity. Thus,

the evolution of speech-based input systems has presented new avenues for the field of data querying and analysis. Several substantial studies published in top database venues have addressed speech-based data querying [30, 29, 19] and DV [32]. While current endeavors in DV have predominantly focused on textual natural language inputs, the potential of speech-based natural language inputs remains largely untapped. Compared to text-based inputs, speech interfaces offer superior user-friendliness and convenience, holding potential in areas such as voice search, hands-free operation, smart home control, and enhancing virtual and augmented reality experiences. In response to the demand for Speech-based DV systems and models, a task named speech-to-visualization (*speech-to-vis*) (Figure 1) has been proposed to generate DV queries from spoken NLQs automatically.

The few existing speech-to-vis work (e.g., Sevi [32]) in the DV domain, simply cascades an Automatic Speech Recognition (ASR) model and a text-to-vis model [18, 31]. However, these systems and approaches suffer from the error propagation problem [30, 29] (i.e., *a small error in the ASR module leads to a much larger error in the following text-to-vis module*) and demand high-quality ASR models for good performance. Designing an end-to-end speech-to-vis system is necessary but challenging. Firstly, the scarcity of speech-to-vis datasets poses a fundamental obstacle, restricting the development and evaluation of end-to-end speech-to-vis models. Secondly, the modality gap between the speech-based query inputs and the textual-based database information presents a tough technical challenge for the neural network.

In response to these challenges, we present the speech-to-vis dataset *SpeechNVBench*, and a novel model named *SpeechVisNet*. To our knowledge, SpeechVisNet is the first end-to-end speech-to-vis model. Our training regimen is structured into three distinct phases: alignment pre-training, weakly supervised data pre-training, and model fine-tuning, each designed to enhance the model's performance and adaptability. Specifically, to alleviate the first challenge, we manually labeled a new speech-to-vis dataset named SpeechNVBench by leveraging the public text-to-vis dataset named NVBench [18], considering factors such as difficulty, length, and domain. Additionally, to further mitigate the situation where the amount of data is insufficient to fully engage model training, we utilized a weakly supervised data pre-training approach to provide preliminary processing before model fine-tuning. To address the second challenge, we designed the sophisticated *SpeechVisNet* model and further employed a methodology, utilizing a teacher-student paradigm for the pre-training step to map the speech-based inputs and textual-based database information into a unified hidden space. SpeechVisNet also employed a DV grammar-aware decoder to generate more rigorous and reliable output. Experiment results show that our model could surpass existing SOTA cascaded baselines.

In summary, our main contributions include:

- We have formalized a novel task named speech-to-vis and have accordingly established the first human-labeled *SpeechNVBench* dataset. As the first manually annotated dataset in this domain, it offers greater possibilities and convenience for following research on this task.
- We introduced SpeechVisNet, a novel model specifically designed for the speech-to-vis task. As the first end-to-end neural network model in its field, this framework

eschews the need for an intermediate text phase, pioneering the capability to directly convert speech into DV.
- We have designed innovative model training methodologies for our designed end-to-end approach, encompassing dual pre-training phases and fine-tuning. The *Alignment Pre-training* aligns the representational spaces of speech and text inputs through a teacher-student framework. Concurrently, the *Weakly Supervised Data Pre-training* enriches the model's training by incorporating labels that bypass manual verification. This synergistic approach of pre-training stages followed by fine-tuning notably boosts the model's capabilities, presenting a valuable blueprint for future reference and in-depth exploration.
- Our SpeechVisNet model has demonstrated exceptional performance in experiments, achieving state-of-the-art (SOTA) results in the task. It outperforms existing baselines by a margin, with an improvement of 9.00% in precision.

2 Task Formulation

In this section, we begin by introducing several preliminary concepts that are instrumental in fostering a deeper understanding of the work that follows, and then we proceed to provide a formal definition of the text-to-vis task.

Natural Language Question. In this paper, an NLQ serves as a comprehensible discourse for humans that describes the desired DV, aligning with people's natural habits of expression and reading. For non-specialists and novices in the field of DV, utilizing NLQs to manipulate and process data is a more user-friendly approach, particularly when the required DV is complex and esoteric.

Visualization Specification. Composing visualization specifications is a crucial step for visualizing data as graphical charts. Declarative visualization languages (DVLs) available in the market are Vega-Lite [26], ggplot2 [36], ZQL [27], ECharts [16], and VizQL [13], each with its grammar. These DVLs detail the visualization's construction, such as chart type, color, size, mapping functions, and properties for marks including canvas size and legends, and thus, determine the visualization specifications. In Figure 1, a specification in Vega-Lite [26] is given, defining attributes such as data path, mark, and encoding.

Data Visualization Query. The DV query concept, proposed to abstract all possible DVLs, allows for the execution of queries on databases to retrieve data, akin to SQL queries. It also provides details necessary for data visualization. In Figure 1, the DV query specifies a "PIE" type chart and defines the data range and aggregation method. This query can be easily converted into a visualization specification in DVLs, which the visualization engine then uses to render the chart. The example in Figure 1 shows this conversion in Vega-Lite, and it is noted that transforming the query for other DVLs, such as ECharts, is straightforward.

Speech-to-Vis Task. Suppose we have a dataset $\mathcal{D}$ with I examples, denoted as $\mathcal{D} = \{\mathbf{d}^1, \cdots, \mathbf{d}^I\}$, where $\mathbf{d}^i$ ($i \in 1, \cdots, I$) refers to the i-th example. Every training example $\mathbf{s}^i$ is structured as $\{x, y, V\}$, with x representing a speech-form NLQ, y denotes its DV query (which can be further executed to obtain the DV chart), and V refers to the schema of the corresponding database needed to execute y. Here, the database schema

V_i contains a set of tables $T = \{t_1, \cdots, t_{M_i}\}$, where M_i represents the count of tables in database schema V_i. For each table t_j ($j \in 1, \cdots, M^i$), it includes a set of columns denoted as $C = \{c_1, \cdots, c_{N_j}\}$, where N_j signifies the count of columns for the table t_j. The goal of the speech-to-vis task is to build a learning-based model that can accurately generate the correct DV query y' from an unseen NLQ-schema pair $\{x', V'\}$. An example is illustrated in Figure 1 to elucidate the task.

3 Speech-to-Vis Dataset

Within this section, we present one of our key contributions: the *SpeechNVBench* dataset. Our presentation will detail its creation methodology, provide an overview of its statistical properties, and elucidate the criteria used for its partitioning.

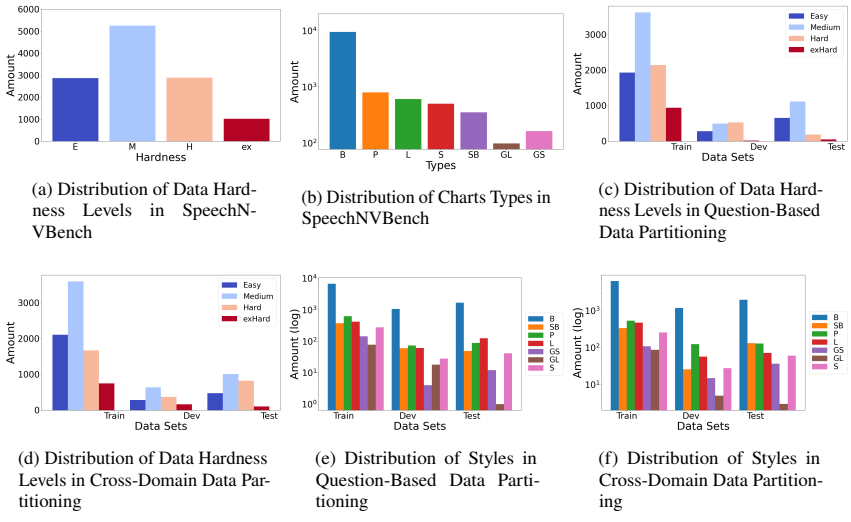
3.1 Dataset Creation Process

To address the data scarcity problem in the speech-to-vis domain, we manually created a dataset named *SpeechNVBench* by refining a subset of 12,000 samples from the NVBench dataset [18]. NVbench introduced a novel synthesizer (nl2sql-to-nl2vis) that transforms the nl-to-SQL (nl2sql) benchmarks into nl-to-vis (nl2vis) benchmarks by analyzing and processing the Abstract Syntax Tree (AST). Furthermore, on this foundation, the data is validated and annotated by experts and crowd workers. On the foundation of this nl2vis benchmark, we meticulously evaluate and filter the data by considering a range of factors including its complexity, length, category, and relevance to specific domains. Specifically, in NVbench, where a one-to-many relationship exists between (NLQs, DVs) pairs, we first guarantee the inclusion of every DV instance within our dataset. Then we endeavor to expand our dataset to a suitable scale while preserving the original benchmark's proportionate distribution of difficulty and types. Subsequently, leveraging the contributions of 32 crowd workers who provided voice recordings, and dedicated 100 hours of meticulous work, we have labeled and constructed the speech-to-vis dataset. In this context, the workers encompass English learners from a wide range of ages, genders, and proficiency levels. They deliver voiceovers in English that, with fluent and largely standard, are characterized by their unique accents, contributing to the rich diversity of the audio data within the dataset. Further details regarding these individuals are available in the appendix.

This process involved a comprehensive assessment of factors such as hardness, length, and domain relevance. Subsequently, we enlisted the assistance of 32 crowd workers, diverse in age and gender, dedicating around 100 hours to annotate and construct this speech-to-vis dataset meticulously.

3.2 Dataset Analysis

Dataset overview. The SpeechNVBench dataset encompasses 153 distinct databases, comprising a total of 780 tables spanning 105 domains. The 153 databases within the dataset contain 780 tables, 4,017 columns, and 1,000,572 rows. So the average number of columns/rows in the 780 tables is 5.15/1,282.78. Ranging from the minimal to the



(a) Distribution of Data Hardness Levels in SpeechNVBench

(b) Distribution of Charts Types in SpeechNVBench

(c) Distribution of Data Hardness Levels in Question-Based Data Partitioning

(d) Distribution of Data Hardness Levels in Cross-Domain Data Partitioning

(e) Distribution of Styles in Question-Based Data Partitioning

(f) Distribution of Styles in Cross-Domain Data Partitioning

Fig. 2: Statistical overview of the complete dataset and its partitioned subsets, including hardness and styles, Categorized by hardness levels Easy(E), Medium(M), Hard(H), and Extra Hard(ex) and Style Types Bar(B), Pie(P), Line(L), Scatter(S), Stacked Bar(SB), Grouping Line (GL), Grouping Scatter (GS).

maximal, the dataset includes a table with just two columns and one row, contrasting with another that spans 48 columns and features 183,978 rows.

(SPEECH, DV) statistics. The WAV2VIS dataset, similar to NVBench [18], encompasses 12,000 pairs of (Speech-form questions, DVs), sharing 7,274 DVs with the NVBench dataset. Each DV in the SpeechNVBench dataset corresponds to one or more Speech-from questions. Among these 12,000 pairs of (Speech-form questions, DVs), they encompass 7 types of charts, further classified into four difficulty levels. The detailed difficulty statistics of this dataset can be found in Figure 2.

3.3 Dataset Partitioning

Two recommended dataset partitioning approaches are available. The first, known as the question-based method, ensures that identical audio inputs are not shared between the test set and the training/validation sets while allowing for the possibility of identical DVs appearing in both. This method is straightforward for model training. The second method termed the cross-domain method, prohibits the presence of the same database in both the test and training/validation sets, requiring stronger model performance and cross-domain transferability. Both approaches are provided within the dataset for users' convenience. We also present specific details regarding the two methodologies of dataset partitioning in Figure 2. Following [31], we use the second approach in our experiments.

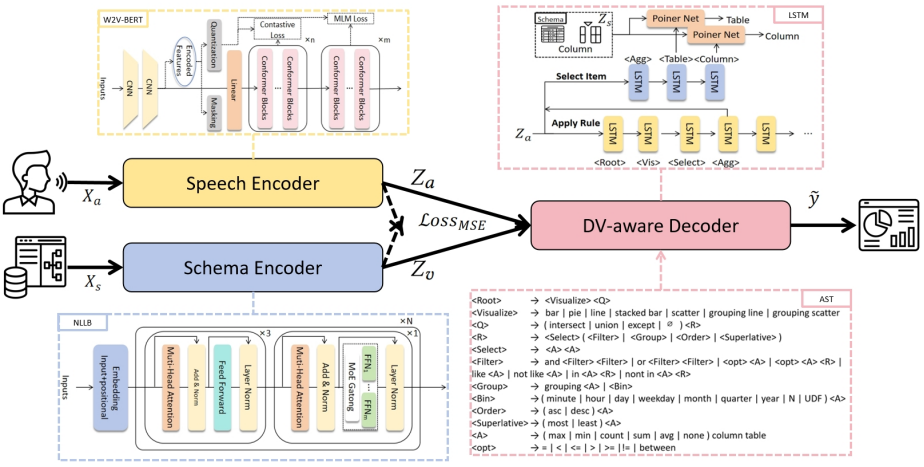


Fig. 3: The network structure of our proposed SpeechVisNet model, which consists of a speech encoder, a schema encoder, and a DV grammar-based decoder to achieve end-to-end speech-to-vis translation. To bridge the modality gap between the speech query (i.e., speech modality) and the database schema (i.e., text modality), a novel pre-training approach is also utilized to align the representations of both speech and text into the same hidden space.

4 Our Proposed Model: SpeechVisNet

In this section, we delve into our SpeechVisNet model, tailored for the speech-to-vis task, beginning with a comprehensive overview of the framework and then proceeding to elucidate the specifics of each constituent part.

4.1 Framework Overview

Due to the significant modality gap between speech NLQ and DV, developing an end-to-end speech-to-vis model presents considerable challenges. Our proposed SpeechVisNet model comprises three main components: a speech encoder, a schema encoder, and a DV-aware decoder. Our speech encoder harnesses the capabilities of the W2V-BERT model [3]. It employs a CNN neural network to process input and leverages the masked language model (MLM) pre-training task of the BERT model to enhance the transformation of speech signals into hidden representations. Simultaneously, information about the tables, columns, and values of the required database is integrated as a natural language input and fed into the schema encoder. We then adopt a pre-training approach inspired by SONAR [9], aligning the representations of both speech and text inputs into the same vector space. Drawing inspiration from existing architectures for Speech-to-SQL tasks [30, 29] and text-to-vis [31] tasks, the DV-aware decoder first predicts an intermediate AST tree using SemQL [12], which then can be translated into a DV query to obtain the final chart. The overall structure of the model is illustrated in Figure 3.

4.2 Schema Encoder

We have adopted the architecture of the NLLB [4] to serve as our schema encoder, which ingests pertinent database information, encompassing table and column names. Specifically, this component is designed with a foundation in transformer mechanisms and is enhanced by the strategic intercalation of a Mixture of Experts (MoE) layer at every three layers, effectively supplanting the traditional Feed-Forward Network (FFN) sublayers. Each MoE layer within the architecture is composed of 128 specialized experts and is paired with a gating mechanism that directs the allocation of tokens. Through the speech encoder, we derived the schema representation Z_v.

4.3 Speech Encoder

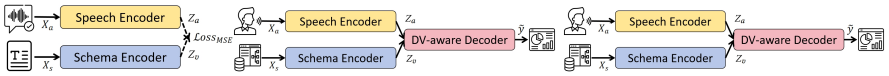
We have constructed a speech encoder architecture inspired by the paradigm of w2v-BERT [3], which adopts conformer layers [11] for constructing the network. As referenced in [11], the conformer architecture—integrated with convolutional neural networks (CNNs) and transformer mechanisms—offers a superior approach to speech modeling. This integration effectively captures the nuanced interplay between the local and global contextual relationships within audio sequences, outperforming a standalone transformer or CNN layers. Upon processing through the speech encoder, we have obtained the speech representation Z_a.

4.4 DV-aware Decoder

Given that DV languages are essentially executable statements grounded in syntactic structures, we opted to design a structured decoder that leverages the syntactic prior knowledge inherent in these languages. In alignment with the established approach in the text-to-SQL field, as delineated in [12], which integrates the SemQL grammar and constructs a corresponding decoder, we have tailored a similar grammar-aware neural architecture. The intricacies of the grammar are illustrated in the lower right corner of Figure 3. Specifically, our DV-aware decoder utilizes an LSTM architecture to generate data visualizations by selecting a sequence of actions represented as $\tilde{y}$. Mathematically, the generation process of a SemQL DV query $\tilde{y}$ can be formalized as follows:

$$p(\tilde{y}|x, V) = \prod_{i=1}^{K} p(act_i|x, V, act_{<i}), \tag{1}$$

where x and V have already been defined in Section 2, and act_i represents an action taken at step i, $act_{<i}$ denotes all actions preceding step i, and K is the total number of actions required to predict $\tilde{y}$. To specify, 'actions' refers to the grammar reasonings in the Rules Application or Schema Selection phase mentioned later. The processes encompassed within the formulation of the equation are delineated into two steps:(i) *Rules Application*: This step involves the application of a production rule to progressively develop the current grammar tree, culminating in the completion of the DV sketch. (ii) *Schema Selection*: This step involves selecting specific column and table elements from the schema to generate the DV query from the sketch.



(a) Alignment Pre-training (b) Weakly Supervised Data Pre-training (c) Model Fine-tuning

Fig. 4: The flowchart of the model training process. Figure (a) corresponds to the alignment pre-training phase, where X_a and X_s signify the speech and textual inputs that correspond to identical content. The outputs of the two encoders are indicated by Z_a and Z_b. Utilizing these, we calculate the MSE Loss, which guides our pre-training process. Furthermore, in Figures (b) and (c), X_a corresponds to the speech query inputs that are crafted in the weakly supervised pre-training stage and the inputs used in the formal training sessions. On the other hand, X_s signifies the text inputs related to the pertinent database information.

Rules Application. The *Rules Application* phase is designed to formulate a context-free grammar tree that underpins the structure of a DV query, following the approach outlined in [12].In this iterative process, we strategically identify the most likely branch at each juncture, leveraging an LSTM-driven framework to navigate the construction. In this iterative process, we discern and opt for the branch with the highest likelihood based on the established route, utilizing an LSTM-based network. Specifically, for each predictive interval i, the LSTM's internal state is updated, contingent upon the previous state $h_{i-1} \in \mathbb{R}^{d_m}$, previous action embedding $act_{i-1} \in \mathbb{R}^{d_a}$ (d_a is the size of the action embedding), previous action type embedding $n_{i-1} \in \mathbb{R}^{d_t}$ (d_t is the size of the action type embedding), and previous context representation of LSTM c_{i-1}. Here, d_a denotes the dimensionality of the action embedding, and d_t denotes the dimensionality of the action type embedding, respectively. Subsequently, the attention context across the encoder's temporal dimensions can be computed, and the production rule can be evaluated using a softmax probability distribution, as presented in Eq. 5.

$$h_i = \texttt{LSTM}([act_{i-1}; n_{i-1}; c_{i-1}], h_{i-1}), \tag{2}$$

$$c_i = \texttt{Softmax}(h_i^T W_a Z_a^T) Z_a, \tag{3}$$

$$u_i = \texttt{tanh}([h_i; c_i] W_u + b_u), \tag{4}$$

$$p(\tilde{y}_i = act_i | x, S, act_{<i}) = \texttt{Softmax}(\texttt{tanh}(u_i^T W_p + b_p)), \tag{5}$$

where $W_a \in \mathbb{R}^{d_m \times d_m}$, $W_u \in \mathbb{R}^{2d_m \times d_m}$ are the learnable weights. When n_a denotes the total number of actions associated with the specified grammar, $W_p \in \mathbb{R}^{d_m \times n_a}$ b_c is trainable bias, and the initial state h_0 is obtained by a max-pooling operation of the speech embedding Z_a.

Schema Selection. To address the task of populating the specific elements within a DV query, we have integrated an LSTM-driven component. The main objective of this module is to decide which item is involved in the text in the condition of the given schema. Items to be decided include tables, columns, and operations like max, min, count, etc. Different from the "rules application" step, the schema exhibits variability across individual cases, with the desired items also lacking a fixed nature. Accordingly,

we have utilized a pointer network, as referenced in Pointer networks [37], to address this challenge. The probability of selecting a schema item is determined through the following computation:

$$p(\tilde{y}_i = act_i | x, V, act_{<i}) = \texttt{Softmax}(u_i^T W_v Z_v^T), \tag{6}$$

where $W_v \in \mathbb{R}^{d_m \times d_m}$ is a learnable weight matrix. Specifically, the selection of table options is confined to those tables that possess the corresponding selected column.

5 Model Training

To bridge the gap between the embedding vector representations of the two modalities and maximize the model's capabilities, we propose a training framework consisting of two pre-training steps followed by fine-tuning: alignment pre-training (Section 5.1), weakly supervised data pre-training (Section 5.2), and model fine-tuning (Section 5.3). We will now delve into a detailed exposition of these components.

5.1 Alignment Pre-training

Following SONAR [9], we use a pre-trained w2v-bert 600 million parameter model to initialize the speech encoders and train them on training sets such as Common Voice 12 ASR [1], Must-C [6], Voxpopuli [38], and Librispeech [22]. After reviewing prior research [9, 8] and conducting a thorough analysis, we opted for Attention-pooling as our ultimate pooling methodology. We chose the Mean Squared Error (MSE) as the objective loss function for our training regimen because it effectively minimizes the average squared difference between the predicted embeddings and the actual ones, thereby ensuring a more accurate representation in the sentence embedding space. This choice aligns with its proven success in prior studies for fostering model performance in multilingual and multimodal contexts [25, 14], as well as in applications involving multilingual speech processing [8]. In particular, we adopted the teacher-student approach, fixing the parameters of the schema encoder, which takes text as input, as the teacher model and minimizing the MSE loss between the two to train the student model, which takes speech as input. The loss function is defined as:

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (h_i^{\text{text}} - h_i^{\text{speech}})^2, \tag{7}$$

where h_i^{text} and h_i^{speech} represent the embeddings of the text and speech inputs at the same timestep, respectively.

To better aggregate information and obtain embeddings of the same size, we employed an attention-pooling mechanism. The output of the attention-pooling layer is calculated as follows:

$$z_i = \sum_{j=1}^{L} \alpha_{ij} h_j, \tag{8}$$

where α_{ij} is the attention weight, computed as:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{L} \exp(e_{ik})}. \tag{9}$$

The attention score e_{ij} is determined by:

$$e_{ij} = \tanh(W_a h_j + b_a), \tag{10}$$

where W_a and b_a are learnable weight matrix and bias vector, respectively. Here, L represents the number of timesteps in the input sequence, i denotes the index of the output embedding, and j denotes the index of the input timestep.

After completing this pre-training phase, we have effectively mapped the two disparate modalities into a unified vector space, establishing a robust groundwork for the following stages of pre-training and the fine-tuning process.

5.2 Weakly Supervised Data Pre-training

Post its preliminary alignment pre-training phase, we generated a significant volume of synthetic data to enhance the model's pre-training process, aiming to unlock its utmost potential for performance. During the weakly supervised data generation phase, we employed Baidu's text-to-speech (TTS) model to provide vocalization for most of the data in NVBench. This approach allowed us to amass a considerable volume of unlabeled weakly supervised data. To elaborate, the synthetic data generated in this phase includes four distinct AI voice profiles, amassing more than 20,000 dubbing outcomes per profile. However, this synthetic dataset intentionally excludes the data from the test and validation sets that are reserved for fine-tuning in the next phase of our process. Given that the training strategies and methodologies applied in this pre-training phase mirror those utilized in the fine-tuning stage, we refrain from repeating them here. For comprehensive insights, one can refer to the section dedicated to fine-tuning for a detailed explanation.

5.3 Model Fine-tuning

Following the pre-training phases previously described, we acquire pre-trained weights for both the speech encoder and the schema encoder. In this subsequent fine-tuning phase, we initially load the pre-existing weights from the pre-training phase, if available. Subsequently, we conduct comprehensive training of the model using speech-to-vis data. This training is aimed at maximizing the log-likelihood of the ground truth action sequences, which are defined as follows:

$$\begin{aligned} \mathcal{L} = \max & \sum_{(x,V,y)\in\mathcal{D}} \sum_{act_i \in ApplyRule} \log p(\tilde{y}_i = act_i | x, V, act_{<i}) \\ & + \sum_{act_i \in SelectSchema} \log p(\tilde{y}_i = act_i | x, V, act_{<i})). \end{aligned} \tag{11}$$

The whole network is trained in an end-to-end style with stochastic gradient descent methods.

6 Experiment

This section presents an in-depth assessment of the performance of our proposed framework, focusing on quantitative metrics. Here, we illustrate the superiority of our framework by juxtaposing its performance with that of several robust baselines.

6.1 Experimental Setup

Datasets We use the same training set of the *SpeechNVBench* dataset, which is detailed in Section 3, to train all the models, tune their parameters with the same validation set, and finally, evaluate the performance on the same testing set. Considering that the data partitioning based on question-based methods presents a lower difficulty level for our model as well as for existing baseline models, and thus does not effectively reflect performance differences. We employ a unified cross-domain approach for dataset partitioning, ensuring that (SPEECH, DV) pairs from the same database do not simultaneously appear in the training/validation sets and the test set. In this scenario, the quantities of the training set, validation set, and test set are 8121, 1442, and 2408 respectively. Concurrently, we strive to ensure an equitable distribution of data hardness and diversity of visualization chart types across all subsets.

Baselines Current approaches addressing the speech-to-vis task are predominantly based on a cascaded methodology, integrating ASR models with text-to-vis models, with a significant lack of end-to-end solutions. In our experiments, we selected four methods as baseline models, comprising both existing cascaded approach models and ASR integrated with existing text-to-vis models. Each model was thoroughly trained under the premise of utilizing the same dataset and adopting an identical dataset partitioning strategy. For our ASR model, we selected the Paraformer [10], a parallel transformer model for non-autoregressive, end-to-end speech recognition. Utilizing the SequenceMatcher class within Python's difflib library, we calculated the degree of similarity between the ASR results and the correct labels, obtaining a notable similarity score of 0.89. It matches the performance of autoregressive models and enhances inference speed through a Lookahead Language Model sampler and minimum word error rate training.

- **ASR-Transformer.** The Transformer architecture has demonstrated its effectiveness in seq2seq tasks. Since the speech-to-vis task falls within the scope of seq2seq tasks, we aim to adopt this most classic model as the fundamental baseline and compare its results with our model's.
- **Sevi (i.e., ASR-ncNet).** Alongside the introduction of the NVBench dataset, Luo *et al.* [18] also proposed ncNet, a model based on the Transformer architecture that leverages templates and incorporates some manual processing techniques. Then they introduced Sevi [32], which integrates ASR and ncNet in a cascading manner.
- **ASR-IRNet.** IRnet [12] is an advanced text-to-SQL model that enhances the traditional direct approach by representing SQL statements in the SemSQL syntax as an AST. We modified IRNet to generate a DV query to adapt to our task.

- **ASR-RGVisNet.** RGVisNet [31] embodies an integration of retrieval and generation modules into a new framework, demonstrating impressive efficacy in addressing the task.
- **SpeechVisNet.** This is the first end-to-end neural model proposed by us that can generate visualizations directly from speech-based questions.

Evaluation Metrics Our experimental validation leverages three well-established metrics from this domain, selected for their capacity to affirm the effectiveness of our innovative model. The first two metrics are applicable to all models within the speech-to-vis domain, while the last one is particularly suited for models that generate intermediate sketches.

- **Exact Match Accuracy.** According to common practice [40], we assess performance through exact match accuracy, where exact match accuracy measures whether the predicted query is equivalent to the gold query as a whole. This stringent metric implies that all elements, such as visualization chart types, x, y, and z-axis, and data with transformations from the according database, must match accurately.
- **Average Time Per Query (TPQ).** This particular metric evaluates the average time expended on inferring a query through various methodologies, thereby reflecting the swiftness with which a model can handle and interpret speech queries.
- **Sketch Match Accuracy.** Our model's decoder generates the final output in two stages: rules application and schema selection. Upon completing the "rules application", we arrive at a sketch that conforms to the DV grammar; think of it as a blueprint for a DV, devoid of specific table column names and values. We assess the alignment of this sketch with the correct one from the labels. The metric, sketch match accuracy, reflects the ratio of accurately generated sketches among the outputs for all queries in the test set. This metric is instrumental in gauging the model's detailed performance, highlighting both its capabilities and areas for improvement.

6.2 Comparison of Accuracy and TPQ

Table 1 displays the accuracy of our proposed model as well as the baselines for the validation and test datasets, from which we can analyze the experimental results.

In our main experiment, the SpeechVisNet model has delivered impressive accuracy, showcasing its competitive edge. Specifically, it surpassed the top-performing baseline model by a notable margin of 9.00% in terms of exact match accuracy on the test dataset. The grammatical intricacies of DV queries prove too sophisticated for the original Transformer architecture, leading to outputs that are vague and devoid of accurate queries. Sevi, which employs templates and some detailed processing, performs relatively better. However, the open-domain dataset's requirement for cross-domain capabilities still limits its performance. The IR-net model and the RGvisnet model, due to their reliance on the accuracy of the ASR component to generate intermediate text-based NLQs, do not perform as impressively as they do in text-to-vis tasks. Overall, our model achieved the highest accuracy and demonstrated the most competitive performance in the main experiments. It does not rely on ASR models and demonstrates robust cross-domain adaptability and flexibility.

Turning our attention to the TPQ, our model demonstrates a markedly shorter processing time for each query compared to other cascaded models, whether they are based on seq2seq or grammar-based approaches. This efficiency stems from the fact that our model operates as an end-to-end, cohesive unit, in contrast to cascaded models which are segmented into distinct components—an ASR module and a text-to-vis module.

Table 1: Performance Comparison

Models	**Acc. (↑)**	**TPQ(s)(↓)**
ASR-IRNet	0.1582	0.3575
ASR-Transformer	0.0	0.6048
ASR-RGVisNet	0.1238	0.3375
Sevi (ASR-ncNet)	0.3095	0.4471
SpeechVisNet(Ours)	**0.3995**	**0.2547**

Table 2: Ablation Experiment

Models	**Sketch Acc.**	**Exact Acc.**
SpeechVisNet	0.7243	0.3995
w/o alignment pre-training	0.2317	0.0
w/o weakly supervised pre-training	0.2552	0.0021

7 Related Work

Our work bridges three key research directions: voice-driven systems, text-to-vis techniques, and speech representation learning.

7.1 Voice-driven Data Querying Systems

Voice-driven systems have evolved significantly in both industry and academia. Early systems like Dragon NaturallySpeaking laid the foundation, while modern platforms (e.g., Siri and Alexa) expanded voice interaction to broader applications. In database querying, EchoQuery [19] pioneered translating specialized voice commands into SQL. Subsequent works like SpeechSQLNet [29] and VoiceQuerySystem [30] generalized this to natural language inputs, enabling intuitive data retrieval for non-experts. For visualization, Sevi [32] introduced a cascaded system combining ASR with a text-to-vis model. However, existing approaches rely on error-prone cascaded pipelines, leaving end-to-end speech-to-vis solutions unexplored—a gap addressed by our work.

7.2 Text-to-Vis Techniques

Text-to-vis research aims to democratize visualization creation. Early systems like Text-to-viz [5] generated infographics from simple textual statistics, while Draco-Learn [20] formalized design constraints. Deep learning approaches emerged with Data2Vis [7], framing visualization as a sequence-to-sequence task. NVBench [18] advanced the field by adapting text-to-SQL benchmarks, enabling transformer-based models like ncNet. Subsequent work integrated retrieval mechanisms (e.g., RGVisNet [31]) and speech inputs (e.g., Sevi [32]), but retained cascaded architectures. Our end-to-end SpeechVisNet eliminates intermediate text conversion, directly mapping speech to visualizations.

7.3 Speech Representation Learning

Self-supervised speech representation learning underpins modern speech systems. CPC [21] pioneered contrastive predictive coding, while Wav2Vec 2.0 [2] integrated transformers for context-aware embeddings. w2v-BERT [3] further unified contrastive and masked language modeling. These advances enabled robust speech encoders critical to our model. By leveraging w2v-BERT's architecture, SpeechVisNet effectively aligns speech signals with textual schema representations, overcoming modality gaps inherent in end-to-end learning.

8 Conclusion and Discussion

In this paper, we introduce a novel speech-driven model for data visualization, along with the associated speech-to-vis task and the SpeechNVBench dataset, aimed at directly converting human-natural language into DV queries. As the pioneering end-to-end solution in its field, our model has been rigorously tested and proven to excel in generating DV queries from spoken inputs. It also exhibits a strong competitive edge when juxtaposed with a multitude of established baselines.

Moving forward, we plan to investigate additional pre-training strategies for systems that operate on voice input. Our ablation experiments revealed a marked decrease in performance when the model was not pre-trained. It is meaningful to examine the causes in-depth, provide a comprehensive analysis, and search for more efficacious pre-training techniques. Furthermore, our SpeechVisNet demonstrates the practicality of a speech-based approach to data visualization, leveraging the methodologies outlined in this research. We are keen to pursue further development in this direction, focusing on crafting more accessible speech-driven systems, particularly for specialized fields such as AI interaction and data analysis fields.

Acknowledgment

We thank the reviewers for their valuable comments. Yuanfeng Song is the corresponding author.

Bibliography

[1] Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., Weber, G.: Common voice: A massively-multilingual speech corpus. In: LREC. pp. 4218–4222. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.520

[2] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[3] Chung, Y.A., Zhang, Y., Han, W., Chiu, C.C., Qin, J., Pang, R., Wu, Y.: W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In: ASRU. pp. 244–250. IEEE (2021)

[4] Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)

[5] Cui, W., Zhang, X., Wang, Y., Huang, H., Chen, B., Fang, L., Zhang, H., Lou, J.G., Zhang, D.: Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. IEEE transactions on visualization and computer graphics **26**(1), 906–916 (2019)

[6] Di Gangi, M.A., Cattoni, R., Bentivogli, L., Negri, M., Turchi, M.: Must-c: a multilingual speech translation corpus. In: NAACL-HLT. pp. 2012–2017. Association for Computational Linguistics (2019)

[7] Dibia, V., Demiralp, Ç.: Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. IEEE computer graphics and applications **39**(5), 33–46 (2019)

[8] Duquenne, P.A., Gong, H., Schwenk, H.: Multimodal and multilingual embeddings for large-scale speech mining. Advances in Neural Information Processing Systems **34**, 15748–15761 (2021)

[9] Duquenne, P.A., Schwenk, H., Sagot, B.: Sonar: sentence-level multimodal and language-agnostic representations. arXiv e-prints pp. arXiv–2308 (2023)

[10] Gao, Z., Zhang, S., McLoughlin, I., Yan, Z.: Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In: Proc. Interspeech 2022. pp. 2063–2067 (2022). https://doi.org/10.21437/Interspeech.2022-9996

[11] Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al.: Conformer: Convolution-augmented transformer for speech recognition. Interspeech 2020 (2020)

[12] Guo, J., Zhan, Z., Gao, Y., Xiao, Y., Lou, J.G., Liu, T., Zhang, D.: Towards complex text-to-sql in cross-domain database with intermediate representation. In: ACL. Association for Computational Linguistics (2019)

[13] Hanrahan, P.: Vizql: a language for query, analysis and visualization. In: SIGMOD. pp. 721–721 (2006)

[14] Heffernan, K., Çelebi, O., Schwenk, H.: Bitext mining using distilled sentence representations for low-resource languages. In: EMNLP. pp. 2101–2112 (2022)
[15] Krommyda, M., Kantere, V.: Visualization systems for linked datasets. In: ICDE. pp. 1790–1793. IEEE (2020)
[16] Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., Zu, M., Chen, W.: Echarts: a declarative framework for rapid construction of web-based visualization. Visual Informatics **2**(2), 136–146 (2018)
[17] Luo, Y., Qin, X., Chai, C., Tang, N., Li, G., Li, W.: Steerable self-driving data visualization. IEEE Transactions on Knowledge and Data Engineering **34**(1), 475–490 (2020)
[18] Luo, Y., Tang, N., Li, G., Chai, C., Li, W., Qin, X.: Synthesizing natural language to visualization (nl2vis) benchmarks from nl2sql benchmarks. In: SIGMOD. pp. 1235–1247 (2021)
[19] Lyons, G., Tran, V., Binnig, C., Cetintemel, U., Kraska, T.: Making the case for query-by-voice with echoquery. In: SIGMOD. pp. 2129–2132 (2016)
[20] Moritz, D., Wang, C., Nelson, G.L., Lin, H., Smith, A.M., Howe, B., Heer, J.: Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. IEEE transactions on visualization and computer graphics **25**(1), 438–448 (2018)
[21] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
[22] Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: ICASSP. pp. 5206–5210. IEEE (2015)
[23] Qian, X., Rossi, R.A., Du, F., Kim, S., Koh, E., Malik, S., Lee, T.Y., Chan, J.: Learning to recommend visualizations from data. In: KDD. pp. 1359–1369 (2021)
[24] Qin, X., Luo, Y., Tang, N., Li, G.: Making data visualization more efficient and effective: a survey. The VLDB Journal **29**(1), 93–117 (2020)
[25] Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: EMNLP. Association for Computational Linguistics (2020)
[26] Satyanarayan, A., Moritz, D., Wongsuphasawat, K., Heer, J.: Vega-lite: A grammar of interactive graphics. IEEE transactions on visualization and computer graphics **23**(1), 341–350 (2016)
[27] Siddiqui, T., Kim, A., Lee, J., Karahalios, K., Parameswaran, A.: Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. Proc. VLDB Endow. **10**(4), 457–468 (nov 2016). https://doi.org/10.14778/3025111.3025126, https://doi.org/10.14778/3025111.3025126
[28] Song, Y., Lu, J., Zhao, X., Wong, R.C.W., Zhang, H.: Demonstration of fevisqa: Free-form question answering over data visualization. In: ICDE. pp. 5417–5420. IEEE (2024)
[29] Song, Y., Wong, R.C.W., Zhao, X.: Speech-to-sql: toward speech-driven sql query generation from natural language question. The VLDB Journal pp. 1–23 (2024)
[30] Song, Y., Wong, R.C.W., Zhao, X., Jiang, D.: Voicequerysystem: A voice-driven database querying system using natural language questions. In: SIGMOD. pp. 2385–2388 (2022)

[31] Song, Y., Zhao, X., Wong, R.C.W., Jiang, D.: Rgvisnet: A hybrid retrieval-generation neural framework towards automatic data visualization generation. In: KDD. pp. 1646–1655 (2022)
[32] Tang, J., Luo, Y., Ouzzani, M., Li, G., Chen, H.: Sevi: Speech-to-visualization through neural machine translation. In: SIGMOD. pp. 2353–2356 (2022)
[33] Tang, N., Wu, E., Li, G.: Towards democratizing relational data visualization. In: Proceedings of the 2019 International Conference on Management of Data. pp. 2025–2030 (2019)
[34] Vartak, M., Huang, S., Siddiqui, T., Madden, S., Parameswaran, A.: Towards visualization recommendation systems. Acm Sigmod Record **45**(4), 34–39 (2017)
[35] Vartak, M., Rahman, S., Madden, S., Parameswaran, A., Polyzotis, N.: Seedb: Efficient data-driven visualization recommendations to support visual analytics. In: VLDB. vol. 8, p. 2182. NIH Public Access (2015)
[36] Villanueva, R.A.M., Chen, Z.J.: ggplot2: elegant graphics for data analysis (2019)
[37] Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. Advances in neural information processing systems **28** (2015)
[38] Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., Dupoux, E.: VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In: ACL-IJCNLP. pp. 993–1003. Association for Computational Linguistics, Online (Aug 2021)
[39] Xie, Y., Luo, Y., Li, G., Tang, N.: Haichart: Human and ai paired visualization system. In: VLDB (2024)
[40] Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., Radev, D.: Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In: EMNLP. Association for Computational Linguistics, Brussels, Belgium (2018)
[41] Yuan, H., Li, G.: A survey of traffic prediction: from spatio-temporal data to intelligent transportation. Data Science and Engineering **6**(1), 63–85 (2021)
[42] Zhang, W., Wang, Y., Song, Y., Wei, V.J., Tian, Y., Qi, Y., Chan, J.H., Wong, R.C.W., Yang, H.: Natural language interfaces for tabular data querying and visualization: A survey. IEEE Transactions on Knowledge and Data Engineering (2024)