
Interactive Cycle Model: The Linkage Combination among Automatic Speech Recognition, Large Language Models, and Smart Glasses

Libo Wang

Nicolaus Copernicus University
Juriija Gagarina 11, 87-100 Toruń, Poland
326360@o365.stud.umk.pl

UCSI University

Taman Connaught, 56000 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia
1002265630@ucsi.university.edu.my

Abstract

This research proposes the interaction loop model "ASR-LLM-Smart Glasses", which model combines automatic speech recognition, large language model and smart glasses to facilitate seamless human-computer interaction. And the methodology of this research involves decomposing the interaction process into different stages and elements. Speech is captured and processed by ASR, then analyzed and interpreted by LLM. The results are then transmitted to smart glasses for display. The feedback loop is complete when the user interacts with the displayed data. Mathematical formulas are used to quantify the performance of the model that revolves around core evaluation points: accuracy, coherence, and latency during ASR speech-to-text conversion. The research results are provided theoretically to test and evaluate the feasibility and performance of the model. Although such human-computer interaction products have not yet appeared in the industry, the performance indicators of this model in enhancing user experience in fields that rely on human-computer interaction have also verified its utility as a technology to promote human-computer interaction. In addition, this research pioneered the idea of integrating cutting-edge technologies such as generative pre-trained Transformer models into unique interaction models, LLM provides raw value through powerful evaluation techniques and innovative use, which provides a new perspective to evaluate and enhanced human-computer interaction.

Keywords: Automatic speech recognition, Large Language Model, Smart glasses, Interaction mechanism

1. Introduction

As the complex and rigorous technology in Natural Language Processing (NLP), the essence of Automatic Speech Recognition (ASR) is regarded as relying on technologies such as machine learning and artificial neural networks in many fields, which convert human spoken language into visual written Basic functions of text to facilitate the development of human-computer interaction (HCI) (Oruh, Viriri & Adegun, 2022; Choutri, Lagha, Meshoul, Batouche, Kacel & Mebarkia, 2022; Padmanabhan & Johnson Premkumar, 2015; Chang, Hung, Wang & Lin, 2011; Huang, Acero, Hon & Reddy, 2001; Lv, Poiesi, Dong, Lloret & Song, 2022). In the initial stage ASR has been able to respond to a limited number of human speech recognition system, and now it has developed into a clear, fluent and accurate complex response system to natural language (Alharbi, Alrazgan, Alrashed, Alnomasi, Almojel, Alharbi, Alturki, Alshehri & Almojel, 2021). Converting human spoken language into written text has long been the fundamental purpose of exploring human-computer interaction (HCI) in this technical field that uses the most natural form of human speech communication to establish an intuitive and effective human-computer interaction model (Bhardwaj, Ben Othman, Kukreja, Belkhier, Bajaj, Goud, Rehman, Shafiq & Hamam, 2022).

Starting from speech capture, the operation process of ASR is closely connected with human spoken instructions as the acoustic speech signal it needs to acquire (Cooke, Barker, Cunningham & Shao, 2006). As the Generative Pre-trained Transformer expands to lattice inputs and continues to change, it stimulates the fields of NLP, LLM and Chatbot to gradually recognize accurate and high-quality speech signals and their clear input as the key to the effectiveness of future human-computer interaction systems. Important links (Huang & Chen, 2019; Jeon, Lee & Choi, 2023). The flow and logic of the design of Generative Pre-trained Transformer 4 (LLM) from few-shot learning is similar to large language model-3.5, Chatlarge language model, BERT and other predecessor models (Liu et al, 2023). It follows an adaptive fine-tuning approach for task-specific adaptations and performs adjustments on small task-specific datasets (Srivastava,

2023; Hu et al, 2023). As the most advanced language model as of the first quarter of 2023, LLM based on Transformer's architectural principles has 175 billion parameters (Baktash & Dawodi, 2023). And for the task of enhancing NLP, LLM has a more complex neural network and a larger training data set (Cheng, Li, Li, Xie, Guo, He & Wu, 2023). Due to the increased model size, it boosts the accuracy performance of the model for human language text understanding compared to previous versions (Singh, SB & Malviya, 2023).

Drawing on descriptions from academic literature, the architecture of LLM is an advanced version of the transformer-based sequence-to-sequence model (Ghojogh & Ghodsi, 2020). Especially its decoder part, which is used as an attention-based for processing sequential data. system (Zhang et al., 2023). The key attention mechanism in this model is called "self-attention", which allows the model to consider the context of the words in the sentence and assign different weights to different words when predicting the next word in the sequence (Biesialska, Biesialska & Rybinski, 2021; Ghojogh & Ghodsi, 2020). While previous Transformer models had both encoders (input processing) and decoders (output processing), LLM and its predecessors are typically decoder-only that only use the decoder components for input processing and output generation (Radford, Wu, Child, Luan, Amodei & Sutskever, 2019).

ASR collects the speech of human users and converts it into the text required by LLM through speech capture, feature extraction, acoustic modeling and language modeling. The accuracy of ASR may be affected by various factors such as the user's English accent, grammatical errors, spoken clarity and fluency problems, and even the background noise level of the user, which may affect the text generated by large language model when performing downstream tasks. (Martín, González-Carrasco, Rodríguez-Fernandez, Souto-Rico, Camacho & Ruiz-Mezcua, 2021). While sequence-to-sequence text generation has been proposed and validated, which attempts to convert incorrect and noisy AS outputs into readable text for humans and downstream tasks (Liao, Shi & Xu, 2022). However, in practice, ASR is disturbed by the above factors during the speech conversion process, which may cause the LLM system to misunderstand and generate inaccurate or incorrect responses.

The integration of ASR technology and LLM has aroused great interest in industry and academic research after OpenAI developed the advanced generative language model Generative Pre-trained Transformer (GPT), which may realize the current human-computer interaction (HCI) paradigm shift (Zhang et al., 2023; Mehrish, Majumder, Bhardwaj & Poria, 2023). Smart glasses are currently known portable wearable electronic computer devices that can provide enhanced visual effects and text information, which include functions such as integrated displays, cameras, microphones, and wireless connections to smartphones or other host devices (Kipper & Rampolla, 2012). The wearer can not only see the real environment from optical display, but also see the virtual content displayed in the display, which is augmented reality (AR) concept (Kurata, Kato, Kurogi, Jung & Endo, 2002). Moreover, products such as smart glasses such as Epson Moverio BT-300, Vuzix Blade, Google Glasses etc. have appeared in the global market, creating conditions for the possible realization of this research (Hou & Bergmann, 2023).

2. Objectives

This research aims to propose the "Interactive Cycle Model" composed of automatic speech recognition, large language model and smart glasses linkage. This model puts forward the idea of enhancing human cognitive ability by wearing human-computer interaction equipment and linking the generated text with large language model. In this linkage model, ASR converts human language into text, which is then processed by LLM, and finally forms a human-computer interaction cycle mechanism by wearing smart glasses that visualize the generated text. This model strives to expand the close connection between the human brain and large language models and achieve human-computer interaction in a way that is less harmful to the human body. It is to predict the possibility of promoting the evolution of human cognition in the direction of omniscience and omnipotence as the large language model model is continuously trained. At the same time, this model provides ideas and reference experience for the development of current smart glasses products that connect with the large language model.

3. ASR-LLM-Smart Glasses

Interactive cycle model is a combination of human instinct and artificial intelligence (Figure 1). It uses the linkage of ASR, large language model and smart glasses to strengthen the close relationship between human users and devices. It uses LLM to generate text and uses ASR to identify human language information to generate what users need. Answers and solutions, and finally visualize them through smart glasses that can both see the real world and possibly see the generated text.

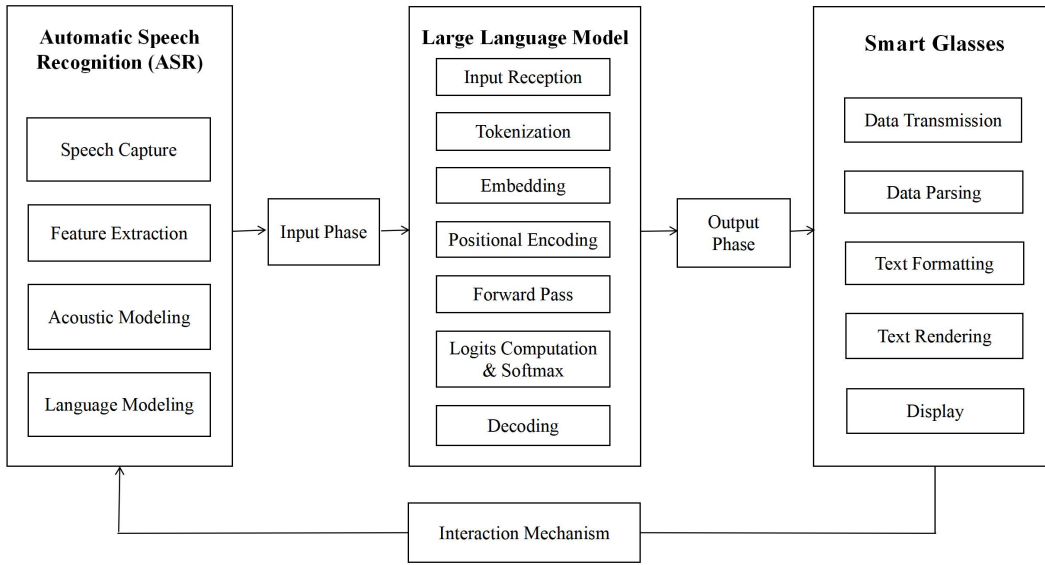


Figure 1 - ASR-LLM-Smart Glasses Linkage Mechanism

4. Literature Review

In order to apply the interactive cycle model to realize that humans can directly obtain massive knowledge and information through language, the following two key problems need to be solved.

4.1 Automatic Speech Recognition (ASR)

Since the core of Automatic Speech Recognition (ASR) technology is based on the integration of linguistics, computer science, and data science, its technical system is dedicated to translating spoken language into written text, enabling audio recording equipment to process conversion users. The speech is text and transmitted to large language model (Ziman, Heusser, Fitzpatrick, Field & Manning, 2018; Anguera, Luque & Gracia, 2014). For users from different regions, ASR needs to fundamentally recognize and understand the complex changes of language in speech that include but not limited to accent, fluency, speech rate, pitch, dialect, and accurately map these information to text (Geng, Xie, Ye, Wang, Li, Hu, Liu, 2022; Eskenazi, 1999; Benzeguiba, Renato de Mori, Deroo, Dupon, Erbes, Jouvett, Fissore, Laface, Mertins, Ris, et al., 2007).

The speech capture stage has a crucial role in ASR that determines how accurately the speech signal can be interpreted and processed by subsequent stages, as a poor quality initial capture can lead to misunderstandings and errors in the final transcription. It is also a major challenge that this research needs to address that captures accurate signals in noisy everyday environments for input to downstream speech recognizers (Haeb-Umbach, Watanabe, Nakatani, Bacchiani, Hoffmeister, Seltzer, Zen & Souden, 2019). In the process of converting sound waves into electrical signals. A description involving sound waves can be expressed mathematically as, and can be described as, an oscillating signal in time:

$$s(t) = A * \sin(2 \pi f t + \phi)$$

$s(t)$ is the sound signal as a function of time; A is the amplitude of the wave (corresponding to the volume of the sound); f is the frequency of the wave (corresponding to the pitch of the sound); ϕ is the phase.

Microphone arrays are speech capture devices used to convert physical speech signals of air pressure changes into electrical signals (Seltzer, 2003). In order to realize that users are not bound by hand-held or head-mounted microphone devices and are easy to carry, device developers may need to consider setting up hands-free speech capture for users (Omologo, Svaizer & Matassoni, 1998). While the hands-free functionality enables the microphone to intelligently capture not only the meaning of speech desired by the user and LLM, but also other unwanted noise-causing signals present in the user's location and environment (Cvijanovic, Sadiq & Srinivasan, 2013). Once the signal is acquired, it is converted to a digital format through a process called Analog-to-Digital Conversion (ADC) that can be done by periodically sampling a continuous analog signal and quantizing the samples into a set of discrete values (Nasiri & Zhang, 2022), and the process is constrained by the Nyquist-Shannon Sampling Theorem (Chen, Yang, Ni, Zhao, Gu, Zhou & Wang, 2016).

As shown in previous literature, the final result obtained in the speech capture stage is represented by the digital time series of the speech signal (Nasereddin & Omari, 2017), which needs to preserve the characteristics of the original speech as much as possible and then extract the features in the subsequent

stages of the ASR process. Characteristics of digital signals (Shrawankar & Thakare, 2013). In ASR, the feature extraction stage is used to convert the raw time-domain speech signal obtained from speech capture into a set of feature vectors that more effectively represent the characteristics of the signal, which is used to separate the different components of speech and improve the efficiency of subsequent stages of ASR is crucial (Han & Wang, 2011). Fourier transform, as a key operation in this step, can be expressed in a mathematical formula:

$$X(k) = \sum_{n=0}^{N-1} [x(n) * e^{(-j*2\pi * k*n/N)}]$$

$X(k)$ is the complex-valued Fourier Transform of the time-domain signal $x(n)$. It's the frequency-domain representation of the signal. \sum is represents summation. (In this context, it means we're summing the given expression for each value of n from 0 to $N-1$). N is the index representing discrete points in the time-domain signal and the total number of samples or data points in the time-domain signal. $x(n)$ is the value of the time-domain signal at time point n . E is Euler's number, a mathematical constant approximately equal to 2.71828. It's the base of the natural logarithm. j is the imaginary unit, which satisfies the equation $j^2 = -1$. In electrical engineering, j is often used instead of i to represent the square root of -1 to avoid confusion with current, denoted by i . 2π comes from the complex exponential representation of a rotating vector in the complex plane, which is the underlying basis of the Fourier Transform. k represents specific frequencies for which we're calculating the Fourier Transform. $e^{(-j*2\pi * k*n/N)}$ represents the complex exponential function, which forms the basis of the Fourier Transform.

Feature extraction can apply a window function to the speech signal, which is usually represented as a Hamming window or Hanning window, which divides the continuous speech signal into smaller overlapping frames, which according to the literature are usually about 20-30 ms long (Ghudasara, Waldekar, Paul & Saha, 2016). This is done based on the quasi-stationary assumption of speech, which suggests that within these small frames, the characteristics of the speech signal do not change significantly. Once the speech signal is framed, a Fast Fourier Transform (FFT) is typically performed on each frame to convert it from the time domain to the frequency domain. FFT is used to extract the spectral features of speech because it captures information about the frequency components of the speech signal and their respective magnitudes. Mel frequency cepstral coefficients (MFCC) or perceptual linear prediction (PLP) coefficients are usually extracted from these frequency domain frames. MFCC or PLP are popular choices because they capture the speech power spectrum in a way that reflects the non-linear human ear perception of different frequencies. Effective results have been achieved in speech recognition systems by using features such as MFCC to fold high-dimensional speech sound waves into low-dimensional codes (Långkvist, Karlsson & Loutfi, 2014). The result of feature extraction is therefore a sequence of feature vectors for each frame of the speech signal (Christ, Braun, Neuffer & Kempa-Liehr, 2018), which provide a more compact and computationally efficient representation of the speech signal, which is then passed to the acoustic modeling stage of the ASR system.

Acoustic modeling is a process step in ASR techniques to represent statistical models of the relationship between acoustic signals and speech units (Tachbelic, Abate & Besacier, 2014). Units of speech include phonemes or phonemes, which are the smallest individual sounds in language for the purpose of creating models that accurately predict the likelihood of a particular phoneme given a particular acoustic signal (Zokirov & Zokirova, 2020). The input to the acoustic model comes from a feature extraction step where the raw speech signal is transformed into a more manageable low-dimensional feature vector (Garg & Sharma, 2016; Gersho & Shoham, 1984). In modern ASR systems, the most commonly used feature type is the Mel-frequency cepstral coefficient (MFCC) that captures the power spectrum of an audio signal in a manner that approximates the response of the human auditory system (Jin, Xu, Xu & Gonzalez, 2013).

Essentially the acoustic modeling task belongs to the pattern recognition problem (Garg & Sharma, 2016) that aims to find the sequence of speech units most likely to produce the observed acoustic features (Chang, Wang, Fang & Qian, 2022). According to the literature, this task is usually accomplished by the statistical model of Hidden Markov Model (HMM), because this model is able to process data sequences and incorporate time dependencies between different parts of the speech signal (Monir, Berbara, Sheikh & Sahidullah, 2021; Tang, 2009). This step involves estimating the probability of a sequence of phonetic feature vectors given a particular word or phoneme sequence, usually using the related formula for forward probability in HMM:

$$\alpha_t(j) = [\sum_{i=1}^N \alpha_{t-1}(i) * a_{ij}] * b_j(O_t)$$

$\alpha_t(j)$ is the forward probability of being in state j at time t , a_{ij} are the state transition probabilities, and $b_j(O_t)$ are the observation likelihoods.

Virtually every phoneme in a language is modeled by its own HMM, whose states represent different parts of the phoneme's pronunciation. For example a simple three-state HMM might have states corresponding to the beginning, middle and end of a phoneme. Transitions between states are governed by transition probabilities, the likelihood of observing a particular eigenvector in each state is given by the state's output probability distribution.

Language modeling is the last key stage of the ASR part of this research. It revolves around the construction of a statistical model that predicts the likelihood of a sequence of words to overcome accuracy challenges due to linguistic uncertainty (Wei, Wang, Wang & Kuo, 2023; Zhang, Cheng, Kumar, Huang, Chen & Mathews, 2022). A language model (LM) aims to capture the probability distribution of a sequence of words in a language, and it is used to predict the probability of the next word given a sequence of words that is based on the history of previous words in the sequence (Li et al, 2022).

N-gram models and neural network-based models are the most commonly used models in ASR systems (Tüske, Schlüter & Ney, 2018; Sivasankaran, Nugraha, Vincent, Morales-Cordovilla, Dalmia, Illina & Liutkus, 2015). The n-gram model approximates the probability of a word given its (n-1) previous words, assuming that the probability of the next word depends only on the previous few words.

$$P(W) = \prod P(w_i | w_1, \dots, w_{i-1})$$

$W = w_1 w_2 \dots w_i$ is a sequence of words in the language; $P(W)$ is the probability of that sequence of words; $P(w_i | w_1, \dots, w_{i-1})$ is the conditional probability of word w_i given the preceding words w_1 through w_{i-1} ; The \prod symbol denotes the product of these conditional probabilities for $i=1$ through n , where n is the length of the sequence.

Binary models (where $n=2$) estimate the probability of a word based on its immediate predecessors, usually assuming that a word's probability depends only on the preceding few words. In a binary model ($n = 2$), the formula simplifies to:

$$P(W) = \prod P(w_i | w_{i-1})$$

This simplifying assumption, known as the Markov assumption, makes the model more tractable.

On the other hand, neural network based LMs are newer and more complex models that can capture long-term dependencies in language (Ji, Feng, Liu, Zhao & Chen, 2022). One example is LM based on recurrent neural networks (RNNs), whose recurrent connections in the network facilitate capturing temporal dependencies in word sequences.

4.2 Large Language Model

LLM is as of now a highly advanced natural language processing (NLP) model based on the Transformer architecture. With the continuous adjustment and optimization of the model design by OpenAI, LLM can capture the context through the self-attention mechanism as the core function, and predict and construct meaningful text output according to the relevance of the context (Koubaa, 2023). LLM is able to tokenize using Byte Pair Encoding (BPE), which alleviates the problems associated with fixed vocabularies by splitting text into sub-word units, allowing dynamic and efficient processing of linguistic phenomena (Zhao et al, 2023 ; Wu, Su, Ma & Liao, 2023). Another key element of LLM is its position embedding layer that assigns high-dimensional vectors to tokens and their respective positions in the input sequence (Wang, Sun, Li, Ouyang Wu, Zhang & Wang, 2023).

As one of the key steps in the LLM operation process in this research, input reception is directly associated with the language modeling stage of the automatic speech recognition (ASR) system by receiving text data for further processing (Patel, Tejani & Talati, 2022). The probability distribution of the word sequence generated by the ASR language model is used as the input of LLM. For LLM, in-context tokenization involves breaking down text data into individual units or tokens, a process handled by Byte Pair Encoding (BPE) that is a technique designed to handle infinite vocabularies in a memory-efficient manner (Petrov, La Malfa, Torr & Bibi, 2023).

The basic functionality of tokenization is dedicated to dividing the input string received from the input sink into smaller units, which are called "tokens". Tokens usually correspond to words or subwords although the granularity may vary (Liu et al., 2021). The tokenization mechanism in LLM mainly uses a process called Byte Pair Encoding (BPE) (Tavabi & Lerman, 2021). BPE is typically a set of individual characters that start with a basic vocabulary and then iteratively merge the most frequently adjacent pairs of symbols in the dataset to build new, longer tokens (Xu & Zhou, 2022; Tacorda, Ignacio, Oco & Roxas, 2017). According to the literature, this method has allowed BPE to make 50,000 tokens such as large language model-2 as a vocabulary of predefined sizes, which can encode any text in the data set, and can also accommodate rare by downgrading to a smaller granularity. and invisible words (Gillioz, Casas, Mugellini & Abou Khaled, 2020; Sennrich, RHHadow & Birch, 2015).

Mathematically, the tokenization process of BPE can be expressed as a function $T(x)$, where x is the input string from the ASR language model:

$$T(x) = \{t_1, t_2, t_3, \dots, t_n\}$$

where $T(x)$ is the token sequence $\{t_1, t_2, t_3, \dots, t_n\}$, and each t_i is a token in the text x . This sequence of tokens is the output of the tokenization process that goes into the next step of embedding.

Embedding is one of the key subsequent steps after tokenization, which plays an important role in the model's ability to understand and generate human language by converting tokens into vectors (Rothman & Gulli, 2022). Essentially, each token in the tokenization stage maps to a high-dimensional vector. The mapping is derived from the learned token embedding matrix $E \in \mathbb{R}^{(d \times |V|)}$, where $|V|$ is the size of the vocabulary and d denotes the embedding dimension. This matrix contains a distinct vector for each token in the vocabulary, and the associated vectors evolve during training to encapsulate semantic and syntactic information about the token.

Mathematically, the embedding operation can be defined as a function $E(t)$, where $t \in T(x)$ is a token in the sequence of tokens produced by the tokenization stage:

$$E(t) = e_t$$

$e_t \in E$ is the corresponding vector in the embedding matrix E of label t .

Since the converter layer consists of a multi-head self-attention mechanism and a feed-forward neural network, the embedding information can be regarded as the basis for the converter layer's processing (Reza, Ferreira, Machado & Tavares, 2022). Self-attention enables complex understanding of syntactic and semantic relationships in text by weighing the importance of each token relative to other tokens in context (Xiao, Hu, Chen, Xue, Chen, Gu & Tang, 2022).

To further ensure linguistic accuracy and context-sensitive text generation, the token vector representations at the embedding stage carry valuable information about the token's meaning and grammatical role in a given linguistic context. The result of this process is a matrix where each row corresponds to a d -dimensional embedding of tokens in the input text, which forms the basis for the positional encoding process (Assylbekov & Takhanov, 2019; Zheng, Ramasinghe & Lucey, 2021).

After the embedding phase, the LLM model enters the positional encoding process that serves as a key step for the model to understand the order or positional information inherent in the input data, which is crucial for language understanding and generating answers (Peng, Li, Zhao & Jin, 2022).

LLM, like other transformer-based models, lacks the inherent ability to recognize sequential information due to the architecture of the model (Chen et al., 2021). To counteract this limitation, LLM introduces positional encodings to token embeddings, ensuring the model understands the relative positions of tokens in the sequence. It allows the model to take into account the temporal dimension of the data, which is crucial for most natural language processing tasks.

Positional encoding process can be expressed by the function $P(i)$, where $i \in \{1, 2, \dots, n\}$ is the position of the token in the sequence:

$$P(i) = p_i$$

p_i is the position encoding vector corresponding to position i .

After obtaining the positional encoding vectors, they are added to the corresponding labeled embedding vectors from the previous embedding stage. Let $E(t)$ be the embedding of token t and $P(i)$ be the positional encoding of position i , token t at position i in the sequence will have a vector representation $V(t, i)$ given by:

$$V(t, i) = E(t) + P(i)$$

The output of this operation is a sequence of vectors, where each vector encapsulates the meaning of the token from the embedding stage and its position in the sequence from the positional encoding stage.

According to the description of BERT in Acheampong, Nunoo-Mensah and Chen (2021) in the study, it can be inferred that LLM from the perspective of principle, after the position encoding stage, the generated vector sequence (encapsulating the semantic representation of the mark and its sequence in the sequence position) into the forward pass phase. At this stage, tokens, embeddings, and positional encodings are passed through a stack of transformer-decoders. Each Transformer's decoder receives input from the self-attention and feed-forward neural network, executes, and so on.

The input to the forward pass can be expressed as $X = [x_1, x_2, \dots, x_n]$, where x_i represents the combined label and position embedding of the label at position i , and n is the total number of labels.

After the self-attention operation, the output is passed through a feed-forward neural network (FFNN). If we denote the output of the self-attention operation as $Z = [z_1, z_2, \dots, z_n]$, then the output of FFNN $Y = [y_1, y_2, \dots, y_n]$ can be calculated as:

$$Y = \text{FFNN}(Z) = \text{ReLU}(Z * W_1 + b_1) * W_2 + b_2$$

Among them, W_1 , b_1 , W_2 , and b_2 are the learning weights and biases of FFNN, and ReLU is the Rectified Linear Unit activation function.

In the Logits Computation & Softmax stage, LLM converts the complex, high-dimensional representation of the tokens sequence into the interpretable probability of the model vocabulary to predict the next token in the sequence (Woodward, Bonnin, Masuda, Varas, Bou-Balust & Riveiro, 2020), which consists of two key operations: Logits calculation and application of softmax function (Deng, Mahadevan & Song, 2023).

Logits are computed via a forward pass on the context-rich embedding $Y = [y_1, y_2, \dots, y_n]$ and the output is projected back into the dimension " $|V|$ " of the model vocabulary V , yielding unnormalized probabilities or "logits". This projection is performed by multiplying the output of the forward pass with the learned weight matrix W_o (of size $|V| \times d_{\text{model}}$, where d_{model} is the dimension of the embedding) and adding a bias term b_o (of size $|V|$) of. This step can be expressed mathematically as:

$$L = Y * W_o + b_o$$

$L = [l_1, l_2, \dots, l_n]$ represents a sequence of logits, where l_i corresponds to the i th marker.

Softmax application: logits $L = [l_1, l_2, \dots, l_n]$ are passed through the softmax function to generate a probability distribution over the vocabulary for each token in the sequence. This function essentially normalizes the logits so that they sum to 1, and the size of each logit determines the probability of the corresponding label.

The softmax function S on logits can be expressed as:

$$S(L) = e^L / \sum(e^L)$$

e^L represents the element-wise exponentiation of the logarithm, and $\sum(e^L)$ ensures that the resulting probabilities sum to 1.

At the completion of the Logits calculation and Softmax phase, the LLM model retains for each token in the input sequence a probability distribution over the entire vocabulary, indicating the likelihood of each possible next token (Kumar, Lu, Gupta, Palepu, Bellamy, Raskar & Beam, 2023).

The forward pass through the decoder can be expressed as:

$$h(l) = \text{LayerNorm}(h(l-1) + \text{MultiHeadAttention}(h(l-1)))$$

In this formula, $h(l)$ is the output l of the th layer; $h(l-1)$ is the output of the previous layer; LayerNorm is layer normalization; MultiHeadAttention is a multi-head self-attention mechanism.

In NLP, the role of the decoder is to convert the output probabilities into a human-readable text format. Similar to BERT and its predecessors, LLM is also a typical decoder-only that uses only the transformer structure of the decoder, which greatly affects the principles underpinning its operation, especially in the decoding stage (Roberts, 2023; Liu, Saleh, Pot, Goodrich, Sepassi, Kaiser & Shazeer, 2018). The encoder processes the input sequence and passes the understanding to the decoder, which then generates the output sequence (Ghojogh & Ghodsi, 2020; Zheng, Zhang & Woodland, 2021). The main advantage of this design is that it allows Transformer models to be trained on large text corpora in an unsupervised manner (Roisenzvit, 2023). It predicts the next token in a sequence based on previous tokens, treating each token in its training corpus as part of an ongoing text (Koubaa, 2023).

Mathematically, the decoding process can be abstracted as:

$$\text{decode}(S(L)) \rightarrow [w_1, w_2, \dots, w_n]$$

$S(L)$ is the softmax function applied to the logits L to obtain a probability distribution over the vocabulary and $[w_1, w_2, \dots, w_n]$ is the sequence of predicted tokens in the output.

When only the decoder LLM performs the decoding operation, it is able to take advantage of the "causal" or "auto-regression" properties of the transformer layer. In other words, each token in the output sequence depends on the preceding tokens, which guides the conversion of the softmax output probabilities into a human-readable sequence. The output generation of LLM relies on the softmax activation function, which is able to transform the high-dimensional vector of each token into a probability distribution in the model vocabulary, so as to select tokens for the output information (Roisenzvit, 2023; Kanai, Fujiwara, Yamanaka & Adachi, 2018).

The self-attention mechanism in the Transformer layer is at the heart of LLM's functionality. It is used to calculate the attention score. The query score Q tokenized by K against key K is computed as the dot product of the two, normalized by the square root of dimension d_k :

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T / \sqrt{d_k})V$$

V is the value vector, K^T is the transpose of the key vector, and the softmax function ensures that the weight sum of all attention scores is 1.

4.3 Smart Glasses

Smart glasses is considered as wearable devices that superimpose digital data onto the user's real-world view (Surti & Mhatre, 2021), and their functionality usually depends on three components: optical system, processing unit, and user interface (Czuszynski, Ruminski, Kocejko & Wtorek, 2015). The processing unit in smart glasses can be thought of as the "smart brain" of the device with management functions (Syberfeldt, Danielsson & Gustavsson, 2017), which is used to process input data from various sensors, execute applications, and generate displays to the user (RajKumar, Arora, Katz, B & Kapila, 2019). The processed text is decoded in LLM and transferred to the smart glasses, text can then be visually displayed to the user through an optical system that superimposes the processed data onto the user's view of the real world.

Smart glasses are considered as wearable devices that superimpose digital data onto the user's real-world view (Surti & Mhatre, 2021), and their functionality usually depends on three components: optical system, processing unit, and user interface (Czuszynski, Ruminski, Kocejko & Wtorek, 2015). The processing unit in smart glasses can be thought of as the "smart brain" of the device with management functions (Syberfeldt, Danielsson & Gustavsson, 2017), which is used to process input data from various sensors, execute applications, and generate displays to the user (RajKumar, Arora, Katz, B & Kapila, 2019). Since the transmission of LLM-generated text occurs over established communication channels, it can be wired (USB) or wireless (Bluetooth or Wi-Fi) (Nam, Park & Kim, 2021). The processed text is decoded in LLM and transferred to the smart glasses, text can then be visually displayed to the user through an optical system that superimposes the processed data onto the user's view of the real world.

Data transfer in context can be achieved through structured procedures. Denote the output of the LLM model as "O_large language model", which represents the text output ready to be displayed. The transfer of O_large language model from the LLM model to the smart glasses can be expressed as a function $T(\bullet)$, thus:

$$T(O_large\ language\ model) \rightarrow O_SG$$

O_SG represents the display output on the smart glasses.

Data parsing is a key process taking place in smart glasses components, which aims to correctly interpret and structure the data received from the previous "data transfer" phase (Novac, 2022; Söldner, Rheinländer, Meyer, Olszowy & Austerjost, 2022). Data parsing breaks down a string of data into smaller tokens, and these components are easier to manage and interpret by the smart glasses' computing system. This process can be represented by a function $P(\)$ acting on the output received from the transmission stage, denoted O_SG.

$$P(O_SG) \rightarrow O_SG_P$$

O_SG_P represents the parsed output, ready to be presented on the display of the smart glasses.

Parsing is usually guided by a set of syntactic rules, which in this case will be dictated by the text formatting and display requirements of the smart glasses user interface.

The concept of "Text Formatting" in the context of smart glasses involves the integration of parsed text data into a format that is both visually appealing and user-readable (Firstenberg & Salas, 2014). After the data is parsed, the formatted text is expressed as $F(\bullet)$, the smart glasses system should express the received information as O_SG_P, and format it for easy display as $F(O_SG_P) \rightarrow O_SG_F$. Text formatting operations are a key factor in ensuring user satisfaction and system usability that involves aspects such as font size, typeface, text alignment, spacing, color, and other visual attributes that can affect readability and the overall user experience (Guo, Daly, Alkan, Mattetti, Cornec & Knijnenburg, 2022).

Text rendering facilitates the implementation of processed and structured data on the interface of smartglasses. Acquired through text formatting, depending on the display technology of the smartglasses (Zhao, Jiang, Chen, Liu, Yang, Xue & Chen, 2022; Huang, 2022; Lee & Hui, 2018). And text rendering may be more proficient in the future to use GPU accelerated operations or dedicated rendering units to accelerate operations to ensure fast, smooth and high-quality output in the intricacies of hardware efficiency.

Display is denoted by $D(\bullet)$ in the smart glasses component and overall interaction cycle model as the final interface between the user and the system, (Rzayev, Woźniak, Dingler & Henze, 2018). Display presents fully rendered text O_SG_R to the user on the visual interface of the smart glasses:

$$D(O_SG_R) \rightarrow O_SG_D$$

The role of the display is to successfully communicate the output of the model to the user in a readable, clear and efficient manner. The text rendering is projected onto the lenses of the smart glasses or an embedded display, forming the final output O_SG_D. Display technologies in smart glasses vary widely that include LED, OLED, or LCOS (Liquid Crystal on Silicon) displays, each with different color reproduction, refresh rate, and power consumption characteristics (Jang, Lee, Kim, Kwak & Park, 2020).

Augmented reality (AR) features can also be incorporated into the display stage, superimposing text output on top of the real-world view (Gabbard, Mehra, & Swan, 2018; Kim, Nussbaum, & Gabbard, 2016).

At the start of the loop, the user issues a command or request $V(t)$, where "t" represents a point in time, and the user's voice input is captured by an automatic speech recognition (ASR) system. The ASR system interprets the utterance and converts it into text, which we denote as $T(V(t))$, where "T" stands for the text transcription function. This transcript is used as input to generate a pretrained Transformer 4 (LLM) model, labeled $G(T(V(t)))$. The "G" stands for the LLM transformation function, which takes the transcribed text and enriches it, producing context-sensitive, coherent text output. Subsequent output from LLM enters the data transfer phase, during which rich text denoted $DT(G(T(V(t))))$ is sent to the smart glasses. The "DT" feature stands for Data Transfer, ensuring that information is passed from LLM to the smart glasses quickly and error-free. The operation chain is denoted as $D(TR(TF(DP(DT(G(T(V(t)))))))$, where DP, TF, TR and D denote data parsing, text formatting, text rendering and Finally, the user receives and processes the displayed information, possibly prompting a new sound input $V(t+1)$, triggering another round of interaction loop.

5. Methodology

Since the "ASR-LLM-Small Glass" model in this research is not a model that already exists and is practiced in the current industry and academia, it is impossible to analyze the model by obtaining data. And the model represents a complex intersection of technologies involving automatic speech recognition (ASR), natural language processing (NLP) using LLM, and smart glasses for information display. Mathematical formulas are suitable for processing and explaining large and complex data sets in the ASR-LLM-Smart Glasses model and obtaining objective and valid results from them. Predictability is the strength of mathematical formulations that extend beyond retrospective analysis by discerning patterns and relationships in data that include predictions of future outcomes or trends in interactive loop models (Liao, Samuel & Krishnamoorthy, 2022).

The use of qualitative approach becomes a crucial research method to validate the recurrent interaction model, since the "ASR-LLM-Smart Glasses" interaction model represents a nascent idea without real-world examples for direct empirical analysis. In contrast, having a mathematical structure as a model component enables the derivation of equations describing the dynamics of interactions independently of reality, and then systematically solves these equations to study the behavior of the model and assess its plausibility (Seshia, Sadigh & Sastry, 2022). Furthermore, mathematical formula allows systematic variation of model parameters, which facilitates the exploration of operational boundaries and conditions under which interaction models can perform optimally (Bubeck et al., 2023).

5.1 Design

The research design of this research is based on a complex recurrent interaction model that contains three key modular components: Automatic Speech Recognition (ASR), Generative Pre-trained Transformer 4 (LLM) and Smart Glasses. Its purpose is to propose ideas and analyze the linkage system operation of the above components through mathematical formula, and to simulate the collaborative operation of the interactive cycle model to realize the interactive cycle mechanism. Starting from the user's voice input, it is processed by ASR, enhanced by LLM and displayed by smart glasses. After the user views the displayed text, there is an audible response, restarting the loop.

The ASR part is subdivided into four steps: speech capture, feature extraction, acoustic modeling, and language modeling. In this model, speech input is recorded and converted into digital data, and unique acoustic features are extracted (Weng, Qin, Tao, Pan, Liu & Li, 2023; Prabhavalkar, Rao, Sainath, Li, Johnson & Jaitly, 2017; Xu, 2022). A string of phonemes is generated when the extracted features are passed to the acoustic model (Dupont, Ris, Deroo & Poitoux, 2005). Phonemes are processed through a language model to generate text strings that represent the most likely interpretation of the user's speech (Kłosowski, 2022).

ASR transcribed text into the LLM module needs to follow the process of the generative pre-training transformer that includes seven steps: input reception, tokenization, embedding, position encoding, forward pass, Logits calculation & Softmax and decoder (Borgeaud et al., 2022; Jaegle et al., 2021; Katz & Belinkov, 2023). The model takes transcribed text as input, tokenizes the text into smaller, manageable units, and creates word embeddings to represent these tokens (Roisenzvit, 2023). The model then assigns positional encodings to preserve the order of words, and the forward pass operation propagates these embeddings through the model. After obtaining the output logits, the Softmax operation converts them into probabilities, and the decoding step finally converts these probabilities back to human-readable text (Sun, Zhang, Huang, Lei, Su, Pan & Cao, 2021).

Refined text data from LLM is data-fed to smart glasses. Smart glasses perform data parsing on the output text to identify the structural components of the text (Grossman, Chen & Fitzmaurice, 2015). Text formatting ensures that the data is styled to fit the glasses display. The text rendering stage converts the formatted text into a form suitable for display, which is then presented to the user in the display step.

5.2 Measurement

Since OpenAI has open sourced the relevant data of LLM, it has credibility, and LLM has been recognized by academic research and the AI industry (Mojadeddi & Rosenberg, 2023). This research defaults to the data provided by OpenAI, and the evaluation and verification of the feasibility and effectiveness of LLM will not appear in this research. Since the accuracy of LLM has been continuously verified and analyzed by users, OpenAI has continuously improved the accuracy of its output text during training (Liu, Ning, Teng, Liu, Zhou & Zhang, 2023; Singh, SB & Malviya, 2023). The technology of displaying text on smart glasses has also been confirmed in industrial applications (Lin et al., 2017). Connect its data transmission capabilities via Bluetooth or Wi-Fi, and the encoded data is transmitted wirelessly to the smart glasses (Jethva, Desai, Verma & Bhagavath, 2020). Therefore, the feasibility of transmitting and displaying text functions has been demonstrated in Google Glasses and previous related literature (Klonoff, 2014; Mishra, 2016; Ward & Helton, 2022). Therefore, the most urgent feasibility that needs to be analyzed and verified in this research is to ensure that the accuracy rate in the process of ASR collecting and converting human speech into text can be quantified and controlled within a reasonable range, which is crucial to user experience (Huh, Park, Lee & Ye, 2023). Because if there is a substantial and serious difference between the user's semantics and the meaning of the text, it may bring inaccuracy to the subsequent language processing stage of LLM, so as to produce wrong or meaningless responses, which may be a fatal blow to the user experience (Wang, Wei, Zhang, Ji & Wang, 2022; Lu, Li & Gong, 2022; Swoboda, Boasen, Léger, Pourchon & Sénécal, 2022). Furthermore, poor recognition accuracy can lead to transcription errors, which in turn lead to inaccurate inputs to the LLM model.

Therefore, this research adopts a qualitative research method and constructs an evaluation system for the relevance, accuracy, delay, and error rate of ASR's conversion of speech into text through formulas in academic literature, which can theoretically promote the effectiveness of the interactive cycle model And is committed to seamless loops to achieve continuous, smooth interaction.

The model is divided into four conceptual steps from exposure to user speech: speech capture, feature extraction, acoustic modeling, and language modeling. Examining automatic speech recognition (ASR) system performance therefore requires a comprehensive evaluation of the different components to reflect their unique contributions to the overall functionality of the ASR system.

Word Error Rate (WER) provides an overall assessment of ASR performance by quantifying ASR systems' quantifying insertions (I), deletions (D), and substitutions (S) compared to human-transcribed reference texts (Ali & Renals, 2018) . The calculation takes into account the total number of words (N) in the reference text. This key evaluation indicator is expressed as:

$$WER = (S + D + I) / N$$

- Voice capture: While there is no clear metric to evaluate the quality of voice capture, an indirect measure, the signal-to-noise ratio (SNR), can be used. A higher SNR indicates superior capture quality.
- Feature extraction: Feature extraction in ASR usually uses Mel-frequency cepstral coefficients (MFCC). While there is no direct measure of feature extraction quality, its impact can be inferred from WER. A high WER may indicate insufficient feature extraction.

Acoustic Modeling: Acoustic model performance can be measured using the Frame Error Rate (FER), a metric that records the proportion of misclassified frames:

$$FER = \text{Incorrectly classified frames} / \text{Total frames}$$

- Language Modeling: Perplexity is a common metric for evaluating language models. A higher language model is indicated by a lower perplexity value. For a test set of N words W1, W2, ..., WN, perplexity is defined as:

$$\text{Perplexity} = (\text{Product from } i=1 \text{ to } N \text{ of } (1/P(W_i|W_{i-1}, \dots, W_1)))^{(1/N)}$$

The combination of the above formulas provides a comprehensive framework for critically evaluating the accuracy of ASR systems.

The coherence of speech-to-text translation of an automatic speech recognition (ASR) system is intricately linked to its ability to maintain logical and contextual continuity in the output text (Narisetty, Tsunoo, Chang, Kashiwagi, Hentschel & Watanabe, 2022; Nyblom, 2022).

Below is a breakdown of potential metrics for the four main components of an ASR system:

- Speech Capture: High quality voice capture ensures that the ASR system has a good starting point for translation. Any misunderstanding at this stage will affect subsequent stages. While there is no specific mathematical formula to measure coherence at this stage, a high signal-to-noise ratio (SNR) will enhance the clarity of captured speech and indirectly improve coherence.

- Feature extraction: In ASR, feature extraction techniques such as Mel-frequency cepstral coefficients (MFCC) are used. Sufficient feature extraction can preserve the essential characteristics of speech, which is necessary for coherence.
- Acoustic Modeling: An acoustic model converts an acoustic signal into a sequence of phonemes or words. The quality of the model will greatly affect the coherence of the output. A low frame error rate (FER) can indicate that the acoustic model is performing well, producing coherent output.

$$FER = \text{misclassified frames} / \text{total frames}$$

- Language Modeling: Coherence is primarily measured at this stage. Perplexity, the inverse probability of the test set, normalized by the number of words, can measure coherence. A lower perplexity score indicates better coherence because the language model can more accurately predict subsequent words in the sequence.

$$\text{Perplexity} = (\text{product from } i=1 \text{ to } N (1/P(W_i|W_{i-1}, \dots, W_1)))^{(1/N)}$$

Given that the nature of coherence is a more qualitative than quantitative feature, the above formulations and assessments provide indirect insight into the coherence of ASR systems.

Delay rate is critical for the seamless functioning of graphs through timely speech-to-text conversion. Delay in ASR can be quantified as the time difference between speech input and its text representation output (Wang, Ren, Qian, Liu, Shi, Qian & Zeng, 2022). Each stage contributes to the total latency and can be rigorously evaluated using specific metrics.

Speech capture: The delay rate of speech capture is highly hardware dependent and can be affected by factors such as microphone quality, signal-to-noise ratio, and network delay rate.

Feature extraction: Feature extraction is a computationally intensive task, so delay rate can be significant, depending on the hardware used. Fetch time "T_{fe}" is often used as a measure of this delay. The exact time depends on factors such as the complexity of the extraction algorithm and the computing power of the hardware.

Acoustic Modeling: Acoustic modeling that translates speech features into a sequence of phonemes or words also contributes to overall latency. This delay "T_{am}" can be calculated by the time it takes to run the acoustic model on the extracted features.

Language Modeling: Language modeling adds a further delay "T_{lm}" as it involves predicting the probability distribution of possible words following a given history of words.

The total delay "D_{total}" can be approximated by summing the individual delays:

$$D_{\text{total}} = T_{\text{fe}} + T_{\text{am}} + T_{\text{lm}}.$$

6. Limitation

After the acquired human speech information is converted into text, the interaction loop model needs to make integration and synchronization of ASR with LLM (Martín, González-Carrasco, Rodríguez-Fernandez, Souto-Rico, Camacho & Ruiz-Mezcua, 2021). After ASR converts the user's voice into text, the output needs to be processed by LLM for natural language understanding and response generation, which means that the linkage process between ASR and LLM technologies and systems is infinitely complex and challenging. seam integration (Geng, Teotia, Tendulkar, Menon & Vondrick, 2023). However, in current practice, real-time synchronization and data exchange between ASR and large language model systems, error handling, and management of potential delays or waiting times may encounter technical obstacles (Alnuhait, Wu & Yu, 2023). Current industry and academia technologies are not yet able to control the accuracy of speech-to-text in interactive loop models within a reasonable range (Liao et al., 2023; Li, Di, Wang, Ouchi & Lu, 2021). And the response speed of the model may be slower than the real-time changes of the text generated by LLM, which may cause high latency to affect user experience (Novac, 2022; Wang et al, 2023; Schaer, Melly, Muller & Widmer, 2016; Meyer, Schlebusch, Spruit, Hellmig & Kasneci, 2021).

7. Summary

This research proposes an interactive cycle model centered on the "ASR-LLM-Smart Glasses" linkage model. It is committed to promoting human-computer interaction and enhancing human capabilities through the combination of voice and vision with the model after the user wears it. The model integrates automatic speech recognition, large language model and a novel human-computer interaction architecture for smart glasses.

In order to verify the objective feasibility of the model in the absence of data due to the absence of relevant physical products, this research introduces the concept of mathematical formula to evaluate and quantify the performance of the model, which includes coherence (C), accuracy (A), Error Rate (E), Delay Rate (D) and Efficiency (Ef) metrics.

This research establishes a solid foundation for the theoretical and practical application of the ASR-LLM-Smart Glasses model. The incorporation of this cutting-edge technology in a seamless interaction model represents a major advance in the field of human-computer interaction.

Reference

- Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, 1-41.
- Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., ... & Almojel, M. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access*, 9, 131858-131876.
- Ali, A., & Renals, S. (2018). Word error rate estimation for speech recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 20-24).
- Alnuhait, D., Wu, Q., & Yu, Z. (2023). FaceChat: An Emotion-Aware Face-to-face Dialogue Framework. *arXiv preprint arXiv:2303.07316*.
- Anguera, X., Luque, J., & Gracia, C. (2014). Audio-to-text alignment for speech recognition with very limited resources. In *Fifteenth Annual Conference of the International Speech Communication Association*.
- Assylbekov, Z., & Takhanov, R. (2019). Context vectors are reflections of word vectors in half the dimensions. *Journal of Artificial Intelligence Research*, 66, 225-242.
- Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Juvet, D., ... & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11), 763-786.
- Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., ... & Hamam, H. (2022). Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, 12(9), 4419.
- Biesialska, M., Biesialska, K., & Rybinski, H. (2021). Leveraging contextual embeddings and self-attention neural networks with bi-attention for sentiment analysis. *Journal of Intelligent Information Systems*, 57(3), 601-626.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning* (pp. 2206-2240). PMLR.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with LLM. *arXiv preprint arXiv:2303.12712*.
- Buhalis, D., & Karatay, N. (2022). Mixed Reality (MR) for generation Z in cultural heritage tourism towards metaverse. In *Information and Communication Technologies in Tourism 2022: Proceedings of the ENTER 2022 eTourism Conference, January 11–14, 2022* (pp. 16-27). Springer International Publishing.
- Chang, S., Wang, X., Fang, T., & Qian, L. (2022). Design and Implementation of Wake-on-Voice and Command Recognition Algorithm. In *2022 14th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* (pp. 1-4). IEEE.
- Chang, Y. S., Hung, S. H., Wang, N. J., & Lin, B. S. (2011). CSR: A Cloud-assisted speech recognition service for personal mobile device. In *2011 International Conference on Parallel Processing* (pp. 305-314). IEEE.
- Chauhan, J., Kaafar, M. A., & Mahanti, A. (2015). Web for Wearables: Lessons learned from Google Glass. *CoRR*.
- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., ... & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34, 15084-15097.
- Chen, Y., Yang, G., Ni, B., Zhao, Z., Gu, X., Zhou, C., & Wang, F. (2016). Development of ground-based ELF/VLF receiver system in Wuhan and its first results. *Advances in Space Research*, 57(9), 1871-1880.
- Cheng, K., Li, Z., Li, C., Xie, R., Guo, Q., He, Y., & Wu, H. (2023). The potential of LLM as an AI-powered virtual assistant for surgeons specialized in joint arthroplasty. *Annals of Biomedical Engineering*, 1-5.
- Choutri, K., Lagha, M., Meshoul, S., Batouche, M., Kacel, Y., & Mebarkia, N. (2022). A Multi-Lingual Speech Recognition-Based Framework to Human-Drone Interaction. *Electronics*, 11(12), 1829.
- Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package). *Neurocomputing*, 307, 72-77.
- Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5), 2421-2424.
- Cvijanovic, N., Sadiq, O., & Srinivasan, S. (2013). Speech enhancement using a remote wireless microphone. *IEEE Transactions on Consumer Electronics*, 59(1), 167-174.
- Czuszynski, K., Ruminski, J., Kocejko, T., & Wtorek, J. (2015). Septic safe interactions with smart glasses in health care. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1604-1607). IEEE.
- Deng, Y., Mahadevan, S., & Song, Z. (2023). Randomized and Deterministic Attention Sparsification Algorithms for Over-parameterized Feature Dimension. *arXiv preprint arXiv:2304.04397*.

- Douillard, A., Ramé, A., Couairon, G., & Cord, M. (2022). Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9285-9295).
- Dupont, S., Ris, C., Deroo, O., & Poitoux, S. (2005). Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents. In *IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 29-34). IEEE.
- Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype.
- Fan, H., Chau, T., Venieris, S. I., Lee, R., Kouris, A., Luk, W., ... & Abdelfattah, M. S. (2022). Adaptable Butterfly Accelerator for Attention-based NNs via Hardware and Algorithm Co-design. In *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)* (pp. 599-615). IEEE.
- Firstenberg, A., & Salas, J. (2014). Designing and developing for Google Glass: Thinking differently for a new platform. "O'Reilly Media, Inc."
- Gabbard, J. L., Mehra, D. G., & Swan, J. E. (2018). Effects of AR display context switching and focal distance switching on human performance. *IEEE transactions on visualization and computer graphics*, 25(6), 2228-2241.
- Garg, A., & Sharma, P. (2016). Survey on acoustic modeling and feature extraction for speech recognition. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 2291-2295). IEEE.
- Geng, M., Xie, X., Ye, Z., Wang, T., Li, G., Hu, S., ... & Meng, H. (2022). Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2597-2611.
- Geng, S., Teotia, R., Tendulkar, P., Menon, S., & Vondrick, C. (2023). Affective Faces for Goal-Driven Dyadic Communication. *arXiv preprint arXiv:2301.10939*.
- Gersho, A., & Shoham, Y. (1984). Hierarchical vector quantization of speech with dynamic codebook allocation. In *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 9, pp. 416-419)*. IEEE.
- Ghodasara, V., Waldekar, S., Paul, D., & Saha, G. (2016). Acoustic scene classification using block-based MFCC features. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*.
- Ghojogh, B., & Ghodsi, A. (2020). Attention mechanism, transformers, BERT, and large language model: Tutorial and survey.
- Gillespie, B. W., & Atlas, A. E. (2003). Strategies for improving audible quality and speech recognition accuracy of reverberant speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. (Vol. 1, pp. I-I). IEEE.
- Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020). Overview of the Transformer-based Models for NLP Tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)* (pp. 179-183). IEEE.
- Grossman, T., Chen, X. A., & Fitzmaurice, G. (2015). Typing on glasses: Adapting text entry to smart eyewear. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 144-152).
- Guo, L., Daly, E. M., Alkan, O., Mattetti, M., Cornec, O., & Knijnenburg, B. (2022). Building trust in interactive machine learning via user contributed interpretable rules. In *27th International Conference on Intelligent User Interfaces* (pp. 537-548).
- Haeb-Umbach, R., Watanabe, S., Nakatani, T., Bacchiani, M., Hoffmeister, B., Seltzer, M. L., ... & Souden, M. (2019). Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal processing magazine*, 36(6), 111-124.
- Han, K., & Wang, D. (2011). An SVM based classification approach to speech separation. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4632-4635). IEEE.
- Heinz, S. (2011). *Mathematical modeling*. Heidelberg: Springer.
- Hou, X., & Bergmann, J. H. (2023). HINNet: Inertial navigation with head-mounted sensors using a neural network. *Engineering Applications of Artificial Intelligence*, 123, 106066.
- Hu, Z., Lan, Y., Wang, L., Xu, W., Lim, E. P., Lee, R. K. W., ... & Poria, S. (2023). LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. *arXiv preprint arXiv:2304.01933*.
- Huang, C. W., & Chen, Y. N. (2019). Adapting pretrained transformer to lattices for spoken language understanding. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 845-852). IEEE.
- Huang, X., Acero, A., Hon, H. W., & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.
- Huang, Y. (2022). AR Shopping List: Exploring the Design Space of Smart Glasses to Allow Real-time Recording with Multiple Input Formats.
- Huh, J., Park, S., Lee, J. E., & Ye, J. C. (2023). Improving Medical Speech-to-Text Accuracy with Vision-Language Pre-training Model. *arXiv preprint arXiv:2303.00091*.

- Jaegle, A., Borgeaud, S., Alayrac, J. B., Doersch, C., Ionescu, C., Ding, D., ... & Carreira, J. (2021). Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795.
- Jamal, N., Shanta, S., Mahmud, F., & Sha'abani, M. N. A. H. (2017). Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review. In AIP Conference Proceedings (Vol. 1883, No. 1, p. 020028). AIP Publishing LLC.
- Jang, H. J., Lee, J. Y., Kim, J., Kwak, J., & Park, J. H. (2020). Progress of display performances: AR, VR, QLED, and OLED. *Journal of Information Display*, 21(1), 1-9.
- Jeon, J., Lee, S., & Choi, S. (2023). A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interactive Learning Environments*, 1-19.
- Jethva, C., Desai, V., Verma, P., & Bhagavath, A. (2020). Smart Glasses and Applications. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, 12(SUP 1), 347-350.
- Ji, P., Feng, Y., Liu, J., Zhao, Z., & Chen, Z. (2022). ASRTest: automated testing for deep-neural-network-driven speech recognition systems. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis* (pp. 189-201).
- Jin, J. S., Xu, C., Xu, M., & Gonzalez, R. (2013). Better than MFCC audio classification features. In *The Era of Interactive Media* (pp. 291-301). Springer New York.
- Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition—a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1, 67.
- Kanai, S., Fujiwara, Y., Yamanaka, Y., & Adachi, S. (2018). Sigsoftmax: Reanalysis of the softmax bottleneck. *Advances in Neural Information Processing Systems*, 31.
- Katz, S., & Belinkov, Y. (2023). Interpreting Transformer's Attention Dynamic Memory and Visualizing the Semantic Information Flow of large language model. arXiv preprint arXiv:2305.13417.
- Klonoff, D. C. (2014). New wearable computers move ahead: Google Glass and smart wigs. *Journal of diabetes science and technology*, 8(1), 3-5.
- Kłosowski, P. (2022). A rule-based grapheme-to-phoneme conversion system. *Applied Sciences*, 12(5), 2758.
- Koubaa, A. (2023). LLM vs. large language model-3.5: A concise showdown.
- Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., & Beam, A. (2023). Conformal Prediction with Large Language Models for Multi-Choice Question Answering. arXiv preprint arXiv:2305.18404.
- Kurata, T., Kato, T., Kourogi, M., Jung, K., & Endo, K. (2002). A Functionally-Distributed Hand Tracking Method for Wearable Visual Interfaces and Its Applications. In *MVA* (pp. 84-89).
- Långkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern recognition letters*, 42, 11-24.
- Lee, L. H., & Hui, P. (2018). Interaction methods for smart glasses: A survey. *IEEE access*, 6, 28712-28732.
- Lee, T., Lee, M. J., Kang, T. G., Jung, S., Kwon, M., Hong, Y., ... & Choi, Y. S. (2021). Adaptable multi-domain language model for transformer asr. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7358-7362). IEEE.
- Li, S., Puig, X., Paxton, C., Du, Y., Wang, C., Fan, L., ... & Zhu, Y. (2022). Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35, 31199-31212.
- Li, W., Di, H., Wang, L., Ouchi, K., & Lu, J. (2021). Boost Transformer with BERT and copying mechanism for ASR error correction. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-6). IEEE.
- Liao, J., Eskimez, S., Lu, L., Shi, Y., Gong, M., Shou, L., ... & Zeng, M. (2023). Improving readability for automatic speech recognition transcription. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5), 1-23.
- Liao, J., Shi, Y., & Xu, Y. (2022). Automatic Speech Recognition Post-Processing for Readability: Task, Dataset and a Two-Stage Pre-Trained Approach. *IEEE Access*, 10, 117053-117066.
- Liao, Z., Samuel, R. D. J., & Krishnamoorthy, S. (2022). Computer vision for facial analysis using human-computer interaction models. *International Journal of Speech Technology*, 25(2), 379-389.
- Lin, Y. C., Liu, J. Y., Wu, Y. C., Ku, P. S., Chen, K., Wu, T. Y., ... & Chen, M. Y. (2017). PeriText+ utilizing peripheral vision for reading text on augmented reality smart glasses. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (pp. 1-3).
- Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of chatlarge language model and LLM. arXiv preprint arXiv:2304.03439.
- Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. arXiv preprint arXiv:1801.10198.
- Liu, X., Yang, B., Liu, D., Zhang, H., Luo, W., Zhang, M., ... & Su, J. (2021). Bridging subword gaps in pretrain-finetune paradigm for natural language generation. arXiv preprint arXiv:2106.06125.

- Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatlarge language model/LLM research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852.
- Lu, L., Li, J., & Gong, Y. (2022). Endpoint detection for streaming end-to-end multi-talker asr. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7312-7316). IEEE.
- Lv, Z., Poiesi, F., Dong, Q., Lloret, J., & Song, H. (2022). Deep Learning for Intelligent Human–Computer Interaction. *Applied Sciences*, 12(22), 11457.
- Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., & Arora, S. (2023). Fine-Tuning Language Models with Just Forward Passes. arXiv preprint arXiv:2305.17333.
- Martin, A., González-Carrasco, I., Rodríguez-Fernandez, V., Souto-Rico, M., Camacho, D., & Ruiz-Mezcua, B. (2021). Deep-Sync: A novel deep learning-based tool for semantic-aware subtitling synchronisation. *Neural Computing and Applications*, 1-15.
- Mehrish, A., Majumder, N., Bhardwaj, R., & Poria, S. (2023). A Review of Deep Learning Techniques for Speech Processing. arXiv preprint arXiv:2305.00359.
- Meyer, J., Schlebusch, T., Spruit, H., Hellmig, J., & Kasneci, E. (2021). A compact low-power gaze gesture sensor based on laser feedback interferometry for smart glasses. In *Digital Optical Technologies 2021* (Vol. 11788, pp. 57-69). SPIE.
- Mishra, S. (2016). Google Glass in medicine—The man with a computer on face. *Indian Heart Journal*, 68(4), 586-587.
- Mojadeddi, Z. M., & Rosenberg, J. (2023). The impact of AI and Chatlarge language model on research reporting. *The New Zealand Medical Journal* (Online), 136(1575), 60-64.
- Monir, N. E., Berbara, Z., Sheikh, I., & Sahidullah, M. (2021). Self-Supervised Learning for Automatic Speech Recognition.
- Nam, M., Park, S., & Kim, D. S. (2021). Intrusion detection method using bi-directional large language model for in-vehicle controller area networks. *IEEE Access*, 9, 124931-124944.
- Narisetty, C., Tsunoo, E., Chang, X., Kashiwagi, Y., Hentschel, M., & Watanabe, S. (2022). Joint speech recognition and audio captioning. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7892-7896). IEEE.
- Nasereddin, H. H., & Omari, A. A. R. (2017). Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation. In *2017 Computing Conference* (pp. 200-207). IEEE.
- Nasiri, H., Li, C., & Zhang, L. (2022). Ultra-Low Power SAR ADC Using Statistical Characteristics of Low-Activity Signals. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 30(9), 1319-1331.
- Nguyen, H., & Santos Jr, E. (2007). An Evaluation of the Accuracy of Capturing User Intent for Information Retrieval. In *IC-AI* (pp. 341-350).
- Novac, P. E. (2022). MicroAI: embedded artificial intelligence for human activity recognition on smart glasses (Doctoral dissertation, Université Côte d'Azur).
- Nyblom, H. (2022). Swedish Language End-to-End Automatic Speech Recognition for Media Monitoring using Deep Learning.
- Omologo, M., Svaizer, P., & Matassoni, M. (1998). Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Communication*, 25(1-3), 75-95.
- Oruh, J., Viriri, S., & Adegun, A. (2022). Long short-term Memory Recurrent neural network for Automatic speech recognition. *IEEE Access*, 10, 30069-30079.
- Padmanabhan, J., & Johnson Premkumar, M. J. (2015). Machine learning in automatic speech recognition: A survey. *IETE Technical Review*, 32(4), 240-251.
- Pan, B., Yang, Y., Li, H., Zhao, Z., Zhuang, Y., Cai, D., & He, X. (2018). Macnet: Transferring knowledge from machine comprehension to sequence-to-sequence models. *Advances in Neural Information Processing Systems*, 31.
- Patel, J., Tejani, R., & Talati, B. (2022). Evaluation Of Performance Of Artificial Intelligence System During Voice Recognition In Social Conversations using NLP. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-4). IEEE.
- Peng, H., Li, G., Zhao, Y., & Jin, Z. (2022). Rethinking Positional Encoding in Tree Transformer for Code Representation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3204-3214).
- Petrov, A., La Malfa, E., Torr, P. H., & Bibi, A. (2023). Language Model Tokenizers Introduce Unfairness Between Languages. arXiv preprint arXiv:2305.15425.
- Prabhavalkar, R., Rao, K., Sainath, T. N., Li, B., Johnson, L., & Jaitly, N. (2017). A Comparison of sequence-to-sequence models for speech recognition. In *Interspeech* (pp. 939-943).
- Rabiner, L. R. (1994). Applications of voice processing to telecommunications. *Proceedings of the IEEE*, 82(2), 199-228.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

- RajKumar, A., Arora, C., Katz, B., & Kapila, V. (2019). Wearable smart glasses for assessment of eye-contact behavior in children with autism. In *Frontiers in Biomedical Devices* (Vol. 41037, p. V001T09A006). American Society of Mechanical Engineers..
- Reza, S., Ferreira, M. C., Machado, J. J. M., & Tavares, J. M. R. (2022). A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. *Expert Systems with Applications*, 202, 117275.
- Roberts, J. (2023). On the Computational Power of Decoder-Only Transformer Language Models. *arXiv preprint arXiv:2305.17026*.
- Rothman, D., & Gulli, A. (2022). *Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and large language model-3*. Packt Publishing Ltd.
- Rzayev, R., Woźniak, P. W., Dingler, T., & Henze, N. (2018). Reading on smart glasses: The effect of text position, presentation type and walking. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-9).
- Saka, A., Taiwo, R., Saka, N., Salami, B., Ajayi, S., Akande, K., & Kazemi, H. (2023). large language model Models in Construction Industry: Opportunities, Limitations, and a Use Case Validation. *arXiv preprint arXiv:2305.18997*.
- Schaer, R., Melly, T., Muller, H., & Widmer, A. (2016). Using smart glasses in medical emergency situations, a qualitative pilot study. In *2016 IEEE Wireless Health (WH)* (pp. 1-5). IEEE.
- Seltzer, M. L. (2003). *Microphone array processing for robust speech recognition* (Doctoral dissertation, Carnegie Mellon University).
- Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- Seshia, S. A., Sadigh, D., & Sastry, S. S. (2022). Toward verified artificial intelligence. *Communications of the ACM*, 65(7), 46-55.
- Shrawankar, U., & Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. *arXiv preprint arXiv:1305.1145*.
- Singh, M., SB, V., & Malviya, N. (2023). Mind meets machine: Unravelling LLM's cognitive psychology. *arXiv preprint arXiv:2303.11436*.
- Sivasankaran, S., Nugraha, A. A., Vincent, E., Morales-Cordovilla, J. A., Dalmia, S., Illina, I., & Liutkus, A. (2015). Robust ASR using neural network based speech enhancement and feature simulation. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 482-489). IEEE.
- Söldner, R., Rheinländer, S., Meyer, T., Olszowy, M., & Austerjost, J. (2022). Human–Device Interaction in the Life Science Laboratory. In *Smart Biolabs of the Future* (pp. 83-113). Cham: Springer International Publishing.
- Srivastava, M. (2023). Exploring the Interplay of Pretraining, Architecture, and Adaptation in Foundation Models.
- Sun, S., Zhang, Z., Huang, B., Lei, P., Su, J., Pan, S., & Cao, J. (2021). Sparse-softmax: A Simpler and Faster Alternative Softmax Transformation. *arXiv preprint arXiv:2112.12433*.
- Surti, P., & Mhatre, P. (2021). *Smart Glasses Technology*.
- Swoboda, D., Boasen, J., Léger, P. M., Pourchon, R., & Sénécal, S. (2022). Comparing the effectiveness of speech and physiological features in explaining emotional responses during voice user interface interactions. *Applied Sciences*, 12(3), 1269.
- Syberfeldt, A., Danielsson, O., & Gustavsson, P. (2017). Augmented reality smart glasses in the smart factory: Product evaluation guidelines and review of available products. *Ieee Access*, 5, 9118-9130.
- Tachbelie, M. Y., Abate, S. T., & Besacier, L. (2014). Using different acoustic, lexical and language modeling units for ASR of an under-resourced language—Amharic. *Speech Communication*, 56, 181-194.
- Tacorda, A. J., Ignacio, M. J., Oco, N., & Roxas, R. E. (2017). Controlling byte pair encoding for neural machine translation. In *2017 International Conference on Asian Language Processing (IALP)* (pp. 168-171). IEEE.
- Tajik, E., & Tajik, F. (2023). A comprehensive Examination of the potential application of Chat large language model in Higher Education Institutions.
- Tang, X. (2009). Hybrid Hidden Markov Model and artificial neural network for automatic speech recognition. In *2009 Pacific-Asia Conference on Circuits, Communications and Systems* (pp. 682-685). IEEE.
- Tavabi, N., & Lerman, K. (2021). Pattern Discovery in Time Series with Byte Pair Encoding. *arXiv preprint arXiv:2106.00614*.
- Tüske, Z., Schlüter, R., & Ney, H. (2018). Investigation on LSTM recurrent n-gram language models for speech recognition. In *Interspeech* (pp. 3358-3362).
- Wang, C. Y., Chu, W. C., Chiu, P. T., Hsiu, M. C., Chiang, Y. H., & Chen, M. Y. (2015). PalmType: Using palms as keyboards for smart glasses. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 153-160).

- Wang, D., Wei, Y., Zhang, K., Ji, D., & Wang, Y. (2022). Automatic speech recognition performance improvement for Mandarin based on optimizing gain control strategy. *Sensors*, 22(8), 3027.
- Wang, J., Scharstein, D., Bapat, A., Blackburn-Matzen, K., Yu, M., Lehman, J., ... & Uyttendaele, M. (2023). A Practical Stereo Depth System for Smart Glasses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21498-21507).
- Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., ... & Wang, G. (2023). large language model-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.
- Wang, W., Ren, S., Qian, Y., Liu, S., Shi, Y., Qian, Y., & Zeng, M. (2022). Optimizing alignment of speech and language latent spaces for end-to-end speech recognition and understanding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7802-7806). IEEE.
- Ward, M. D., & Helton, W. S. (2022). Dual-task interference while receiving information on a head mounted display and manual tracking with and without auditory warnings. *Applied ergonomics*, 101, 103713.
- Wei, C., Wang, Y. C., Wang, B., & Kuo, C. C. J. (2023). An overview on language models: Recent developments and outlook. *arXiv preprint arXiv:2303.05759*.
- Weng, Y., Miryala, S. S., Khatri, C., Wang, R., Zheng, H., Molino, P., ... & Tur, G. (2020). Joint contextual modeling for asr correction and language understanding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6349-6353). IEEE.
- Weng, Z., Qin, Z., Tao, X., Pan, C., Liu, G., & Li, G. Y. (2023). Deep learning enabled semantic communications with speech recognition and synthesis. *IEEE Transactions on Wireless Communications*.
- Woodward, A., Bonnin, C., Masuda, I., Varas, D., Bou-Balust, E., & Riveiro, J. C. (2020). Confidence Measures in Encoder-Decoder Models for Speech Recognition. In *INTERSPEECH* (pp. 611-615).
- Xia, Y., Khamis, M., Fernandez, F. A., Heidari, H., Butt, H., Ahmed, Z., ... & Ghannam, R. (2022). State-of-the-Art in Smart Contact Lenses for Human–Machine Interaction. *IEEE Transactions on Human-Machine Systems*.
- Xiao, L., Hu, X., Chen, Y., Xue, Y., Chen, B., Gu, D., & Tang, B. (2022). Multi-head self-attention based gated graph convolutional networks for aspect-based sentiment classification. *Multimedia Tools and Applications*, 1-20.
- Xu, H. (2022). Intelligent automobile auxiliary propagation system based on speech recognition and AI driven feature extraction techniques. *International Journal of Speech Technology*, 1-13.
- Xu, T., & Zhou, P. (2022). Feature Extraction for Payload Classification: A Byte Pair Encoding Algorithm. In *2022 IEEE 8th International Conference on Computer and Communications (ICCC)* (pp. 1-5). IEEE.
- Yang, C. C., Fan, W. C., Yang, C. F., & Wang, Y. C. F. (2022). Cross-Modal Mutual Learning for Audio-Visual Speech Recognition and Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 3, pp. 3036-3044).
- Yu, S. J., Sun, J. C. Y., & Chen, O. T. C. (2019). Effect of AR-based online wearable guides on university students' situational interest and learning performance. *Universal Access in the Information Society*, 18, 287-299.
- Zhang, C., Zhang, C., Li, C., Qiao, Y., Zheng, S., Dam, S. K., ... & Hong, C. S. (2023). One small step for generative ai, one giant leap for agi: A complete survey on chatlarge language model in aigc era. *arXiv preprint arXiv:2304.06488*.
- Zhang, H., Cheng, Y. C., Kumar, S., Huang, W. R., Chen, M., & Mathews, R. (2022). Capitalization Normalization for Language Modeling with an Accurate and Efficient Hierarchical RNN Model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6097-6101). IEEE.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Zhao, Y., Jiang, J., Chen, Y., Liu, R., Yang, Y., Xue, X., & Chen, S. (2022). Metaverse: Perspectives from graphics, interactions and visualization. *Visual Informatics*.
- Zheng, J., Ramasinghe, S., & Lucey, S. (2021). Rethinking positional encoding. *arXiv preprint arXiv:2107.02561*.
- Zheng, X., Zhang, C., & Woodland, P. C. (2021). Adapting large language model, large language model-2 and BERT language models for speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 162-168). IEEE.
- Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., & Manning, J. R. (2018). Is automatic speech-to-text transcription ready for use in psychological experiments?. *Behavior research methods*, 50, 2597-2605.
- Zokirov, M. T., & Zokirova, S. M. (2020). Contrastive analysis at the phonetic level. *Academic Leadership (Online Journal)*, 21(05), 163-169.