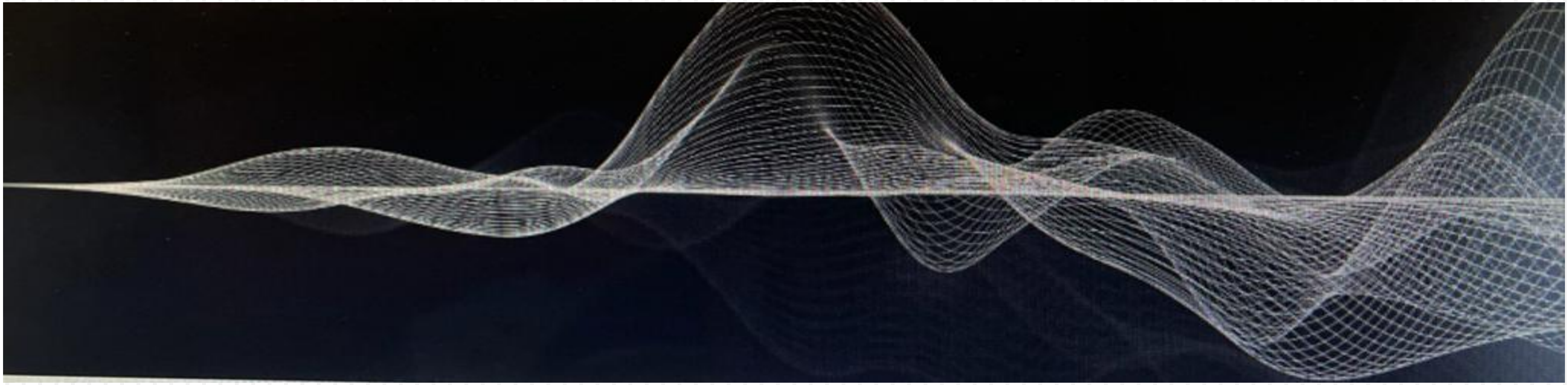
# AI-based Speech Recognition Technology for Asian & Austronesian Languages



Hary Gunarto

**Ritsumeikan Asia Pacific University**

**AP Conference, Japan 30 Nov. 2024**

# ABSTRACT

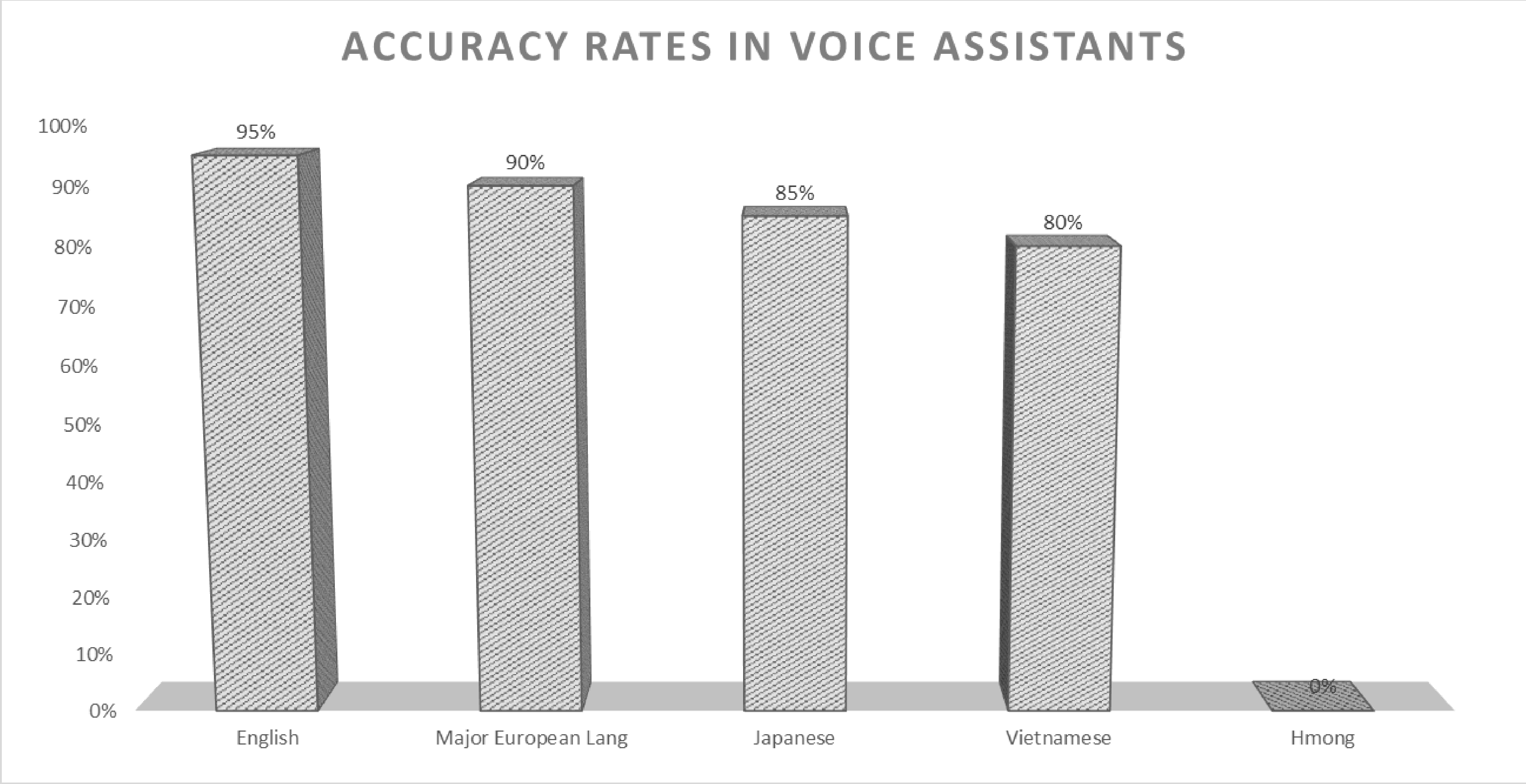
Recent progress in speech recognition systems is changing toward advancement of linguistic processing, particularly for Asian and Austronesian languages. These languages, characterized by many and various phonetic and tonal variations, present distinctive challenges for voice and speech recognition systems. Recent research has focused on developing advanced algorithms that are able to effectively handle these complexities.

Methodology and innovations in AI, deep learning and ChatGPT.4o, particularly the utilization of computational Recurrent Neural Networks (RNN) and Edge computing, have significantly enhanced the accuracy of voice recognition in these languages. Moreover, researchers are implementing large-scale multilingual datasets and learning techniques to train models that are more adaptable to the specific linguistic features in these languages.

Furthermore, developments in acoustic modeling and the integration of phonetic and intonation features have further improved system performance. Efforts are also being made to address and quickly respond to urgent messages in dangerous situations and requests. These developments are not only increasing the precision but also expanding their applications in various fields such as home, healthcare, education, transportation and telecommunications.

In conclusion, recent computational AI-based voice recognition is becoming an essential tool for linguistic diversity in the Asian and Austronesian regions, promoting greater accessibility and use. It is time now to express our language through math—sentences with variables, paragraphs with equations, and language context with formulas.

# LANGUAGE PROCESSING ACCURACY



# Asian & Austronesian Lang’s Tones

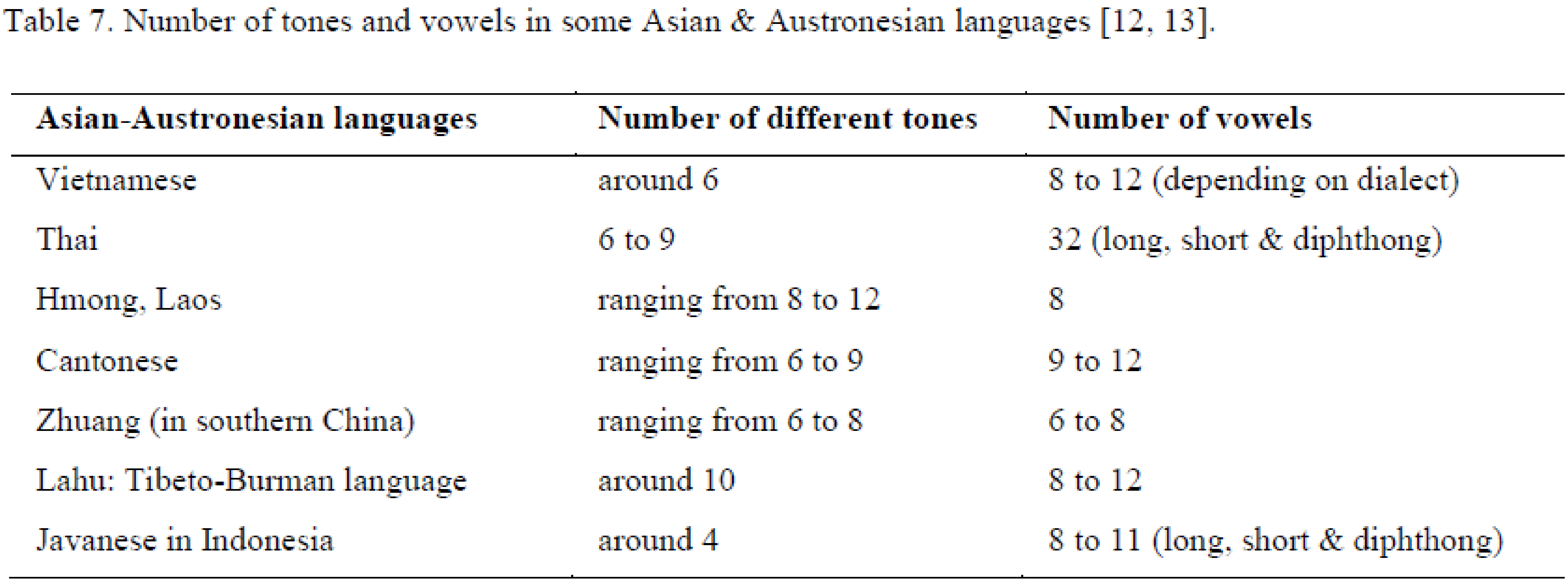


Table 7. Number of tones and vowels in some Asian & Austronesian languages [12, 13].

| Asian-Austronesian languages | Number of different tones | Number of vowels |
|---|---|---|
| Vietnamese | around 6 | 8 to 12 (depending on dialect) |
| Thai | 6 to 9 | 32 (long, short & diphthong) |
| Hmong, Laos | ranging from 8 to 12 | 8 |
| Cantonese | ranging from 6 to 9 | 9 to 12 |
| Zhuang (in southern China) | ranging from 6 to 8 | 6 to 8 |
| Lahu: Tibeto-Burman language | around 10 | 8 to 12 |
| Javanese in Indonesia | around 4 | 8 to 11 (long, short & diphthong) |

# Phonetic & Homophones Problems

Homophones are words that sound the same but have different meanings, such as “two” and “too”, “write” and “right.” Ambiguities in speech, such as sentence-level ambiguities of words with multiple interpretations, can lead to misinterpretations in voice recognition systems. In Asian languages especially Japanese, there are a lot of such word ambiguities, e.g.,

- kiku ” 聞く ” (to hear or listen),
- kiku “訊く ” (to ask or query) and
- kiku “効く ” (to be effective), and other 5 ‘kiku’s.

Homophones 玩 “wán” (to play) and 晚 “wǎn” (late) is another example in Chinese Mandarin.

Besides Japanese and Mandarin, Vietnamese has a vast number of homonyms due to its relatively many sounds inventory (syllables pronounced with different tones), such as in Cầm (to hold) and Cấm (forbidden) or in Mắt (eye) and Mạt (end).

# AI-based Voice Recognition System Software

## for Asian & Austronesian Languages

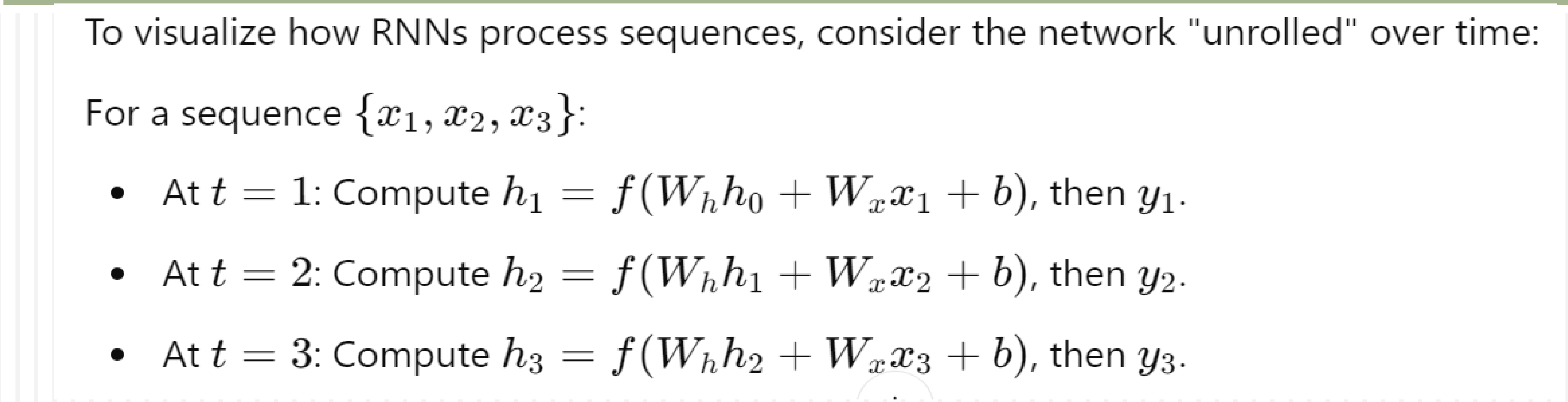
1. **Google's Multilingual Neural Machine Translation**
2. **Chinese** iFLYTEK AI Speech Recognition
3. **Mozilla DeepSpeech:** Thai, Vietnamese, and Filipino
4. **Baidu Deep Speech:** Mandarin and other Chinese dialects
5. **South Asian** OpenNMT (Open Neural Machine Translation)
6. **Wav2Vec 2.0 by Facebook AI:** Tagalog, Cebuano, Javanese
7. others.

These tools provide pre-trained models and offer flexibility to use custom models on phonetic and tonal datasets.

# Several language models & Architecture

| **Software** | **Model Type(s)** | **Main Architecture** |
|---|---|---|
| **Google GNMT** | Transformer | Attention-based transformers |
| **Mozilla DeepSpeech** | RNN, LSTM | Sequence-to-sequence with CTC |
| **iFLYTEK AI** | Acoustic + DNN | Custom deep learning models |
| **Wav2Vec 2.0** | CNN + Transformer | Self-supervised learning, transformers |
| **Kaldi** | GMM-HMM, LSTM, Transformer (flexible) | Hybrid models, sequence learning |
| **OpenNMT** | Transformer | Attention-based transformers |
| **Baidu Deep Speech** | RNN, later transformers and CNN | Sequence-to-sequence with CTC |

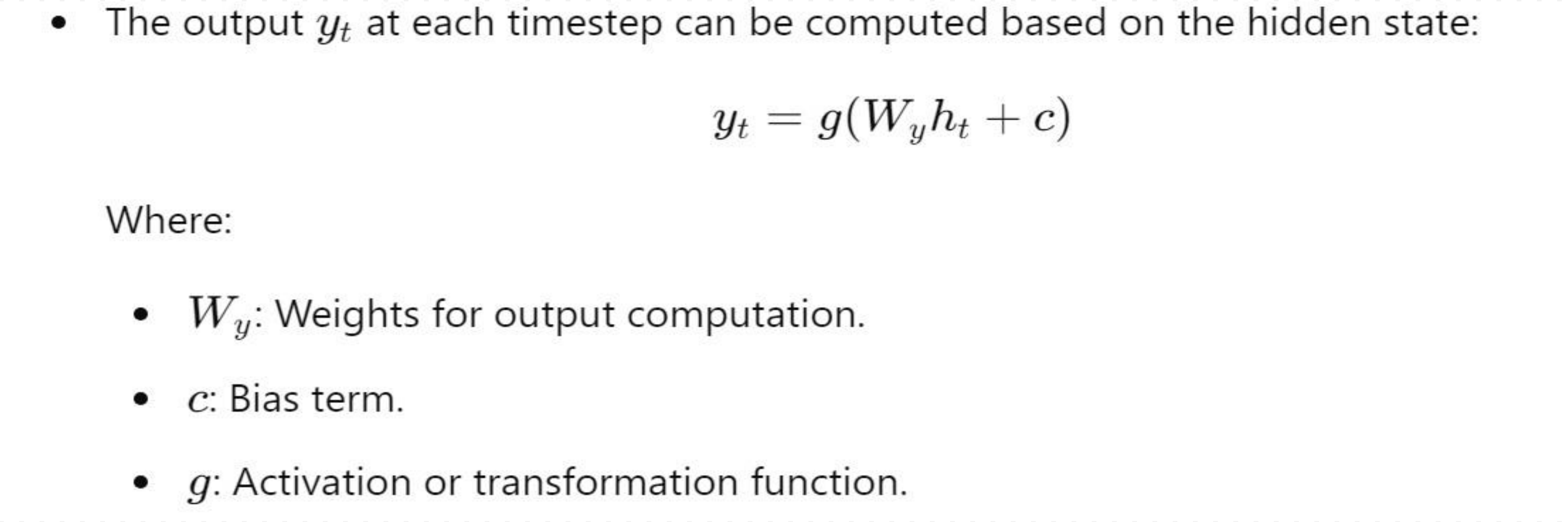
# Computational Language RNN Model



To visualize how RNNs process sequences, consider the network "unrolled" over time:

For a sequence $\{x_1, x_2, x_3\}$:

- At $t = 1$: Compute $h_1 = f(W_h h_0 + W_x x_1 + b)$, then $y_1$.
- At $t = 2$: Compute $h_2 = f(W_h h_1 + W_x x_2 + b)$, then $y_2$.
- At $t = 3$: Compute $h_3 = f(W_h h_2 + W_x x_3 + b)$, then $y_3$.



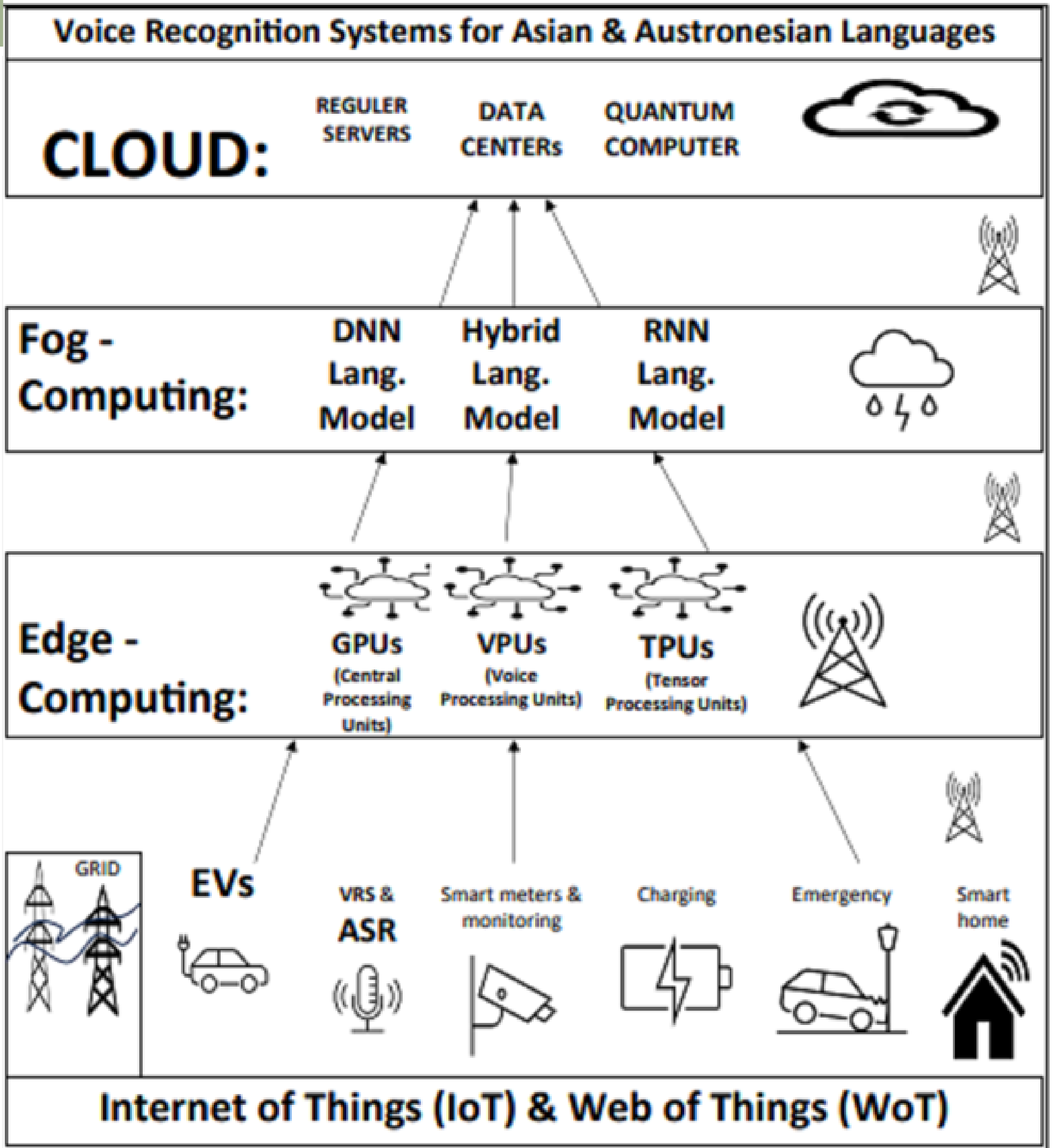
- The output $y_t$ at each timestep can be computed based on the hidden state:

$$y_t = g(W_y h_t + c)$$

Where:

- $W_y$: Weights for output computation.
- $c$: Bias term.
- $g$: Activation or transformation function.

# Lang. Processing Network Architecture


Voice Recognition Systems for Asian & Austronesian Languages
CLOUD:
REGULER SERVERS
DATA CENTERs
QUANTUM COMPUTER
Fog - Computing:
DNN Lang. Model
Hybrid Lang. Model
RNN Lang. Model
Edge - Computing:
GPUs (Central Processing Units)
VPUs (Voice Processing Units)
TPUs (Tensor Processing Units)
GRID
EVs
VRS & ASR
Smart meters & monitoring
Charging
Emergency
Smart home
Internet of Things (IoT) & Web of Things (WoT)

# Challenges & future developments

- Tonal & Phonetic Complexity: in Mandarin, "ma" has four tones with different meaning (mother, hemp, horse, or scold: mā, má, mǎ, mà, 媽、麻、馬、罵).
- Dialectal Variations and Localized Accents
- Script and Transliteration Complexity
- Language Mixing
- Contextual Understanding and politeness levels.

# Thank you

"Mathematics and computational languages are among the most thrilling and transformative tools for AI-driven expression and voice communication ever created by humanity."

## References:

* Gunarto Hary, Applications of AI-empowered electric vehicles for voice recognition in Asian and Austronesian languages, *on: Artificial Intelligence-Empowered Modern Electric Vehicles in Smart Grid Systems, Elsevier, May 2024, ISBN: 978-0443238147, pp 81-112.*

* Brunelle M and Kirby J. *Tone and Phonation in Southeast Asian Languages*. First published: 03 April 2016 https://doi.org/10.1111/lnc3.12182.

* Bernard C. *The World's Major Languages*. Routledge, Third edition, 2018, New York and London.