Vietnam National University, Ho Chi Minh City
University of Information Technology

Faculty of Computer Science

SPEECH INFORMATION PROCESSING

EXERCISE NO. 5 REPORT

Speech Synthesis

Group: 3 – Ngũ Cốc
Class: CS410.P21

Date: April 9th, 2025
Instructor: MSc. Nguyễn Thành Luân

Contents

1. Diphone Synthesis and Unit Selection

1.1 Diphone Synthesis

1.1.1 What is a Diphone?

First, let us understand what a “phone” is, which is the uttermost unit of speech, i.e. the phoneme; then, “di” is a latin-originated prefix which roughly translate to “to be made of two”. We can now safely presume that a diphone is a combination of two “phones”, a more composed unit of utterance that, by nature, encompasses the transition between two adjacent phonemes. Our presumption is easily justified with Collins Dictionary definition.

1.1.2 Synthesis with Diphones

Diphone Synthesis is one of the most fundamental techniques used for creating a synthetic voice from recordings or samples of a particular speaker; it captures a wide range of the acoustic quality of an individual, within some limits. Opting for a diphone-based synthesis approach allows for a more natural and expressive voice, as it takes into account the subtle transitions between phonemes. More specifically, diphone synthesis shift the focus point from the phoneme, which is the stable region of a phonetic realization, to the transition phase between two adjacent instances of phones, known for its inherent variability and complexity, resulting in a more rigid challenge in modeling. The diphone, then, cuts the units at the points of relative stability, rather than at the volatile phone-phone transition, where so-called coarticulatory effects appear [1].

1.2 Unit Selection

1.2.1 What seems to be a Unit?

Differs from Diphone Synthesis, Unit Selection does not synthesize speech by concatenating just small recordings of phonemes, but rather by selecting and combining pre-recorded sounds of various lengths including phonemes, syllables, morphemems, words, phrases or sentences that best match the desired speech characteristics; these pre-recorded items are the units of Unit Selection [2].

1.2.2 How does it work?

A large speech corpus is first created by recording a speaker and then carefully segmented into smaller units, such as phonemes, syllables, or words. Each segmented unit is analyzed and labeled with key acoustic properties like pitch, duration, and spectral features. When synthesizing new speech, the system analyzes the target text to determine the desired speech characteristics. It then searches the indexed database to find candidate units that best match these properties, balancing two factors: the target cost, which measures how well the unit matches the desired sound, and the concatenation cost, which evaluates how smoothly the unit can join with its neighbors. The optimal sequence of units is selected by minimizing these costs, and they are concatenated to form the final natural-sounding speech output [2].

2. Target Cost and Join Cost

2.1 Target Cost

In concatenative speech synthesis (CSS), the target cost quantifies how well a candidate speech unit matches the desired attributes of the target utterance. This evaluation involves comparing the unit's phonetic and prosodic features—including phoneme identity, duration, pitch, and contextual nuances—to those specified by the target. Each individual feature is assessed to produce subcosts, which are then weighted according to their importance in achieving natural-sounding speech. By combining these subcosts into an overall target cost, the system is able to select speech units that closely align with the intended speech characteristics, leading to increased intelligibility and natural quality in the synthesized output [3].

2.2 Join Cost

The join cost, formerly referred to as the concatenation cost, evaluates the acoustic compatibility between adjacent speech units to ensure smooth and natural transitions in synthesized speech. This cost function focuses on the continuity at the boundaries of the units by analyzing various acoustic features such as spectral similarity, pitch consistency, and energy distribution. Each of these factors is individually assessed, and the resulting subcosts are aggregated into a weighted sum that represents the overall join cost. Minimizing this cost is crucial to avoid perceptible discontinuities, thereby producing fluent, cohesive, and realistic synthesized speech [3].

Bibliography

[1] K. Lenzo and A. Black, "Diphone collection and synthesis," 11 2003.

[2] S. N. Kayte, M. Mal, and C. Kayte, "A review of unit selection speech synthesis," vol. 5, p. 5, 11 2015.

[3] T. Bäckström, O. Räsänen, A. Zewoudie, P. P. Zarazaga, L. Koivusalo, S. Das, E. G. Mellado, M. B. Mansali, and D. Ramos, *Introduction to Speech Processing: 2nd Edition.* Zenodo, Jul. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6821775