DEVELOPMENT AND EVALUATION OF JAPANESE TEXT-TO-SPEECH MIDDLEWARE FOR 32-BIT MICROCONTROLLERS

Nobuyuki Nishizawa, Tomohiro Obara and Gen Hattori (KDDI Research, Inc., Japan)

SUMMARY

HMM-based standalone Japanese text-to-speech middleware for microcontrollers have been developed:
Key techniques:
- Compression of the morphological dictionary by LOUDS
- Sinusoidal synthesis on a subband coding system
- Parameter generation with fixed-point arithmetic
- Pipelined processing

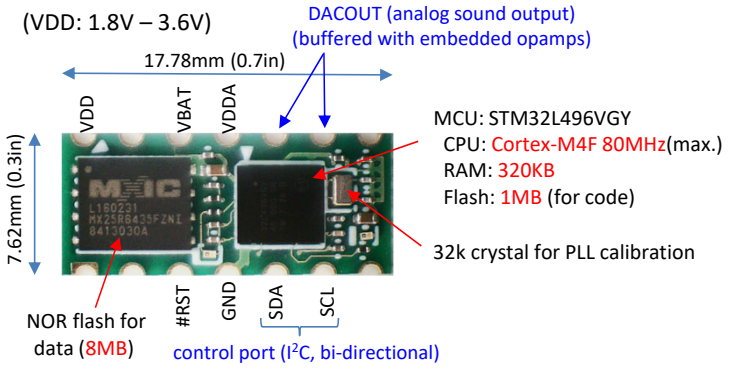
Realtime processing is achieved with consumption current < 15mA

Targets of our text-to-speech (TTS) middleware

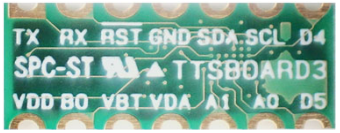
- Wearable or compact IoT devices with no or narrowband connectivity (e.g. <1kbps)
- Low-power consumption is often demanded.

Miniature MCU board as a testing platform

MCU (microcontroller unit):
One-package device including CPU cores, clocks, SRAM, flash memory, timers, GPIOs, digital-to-analog converters (DACs), ….

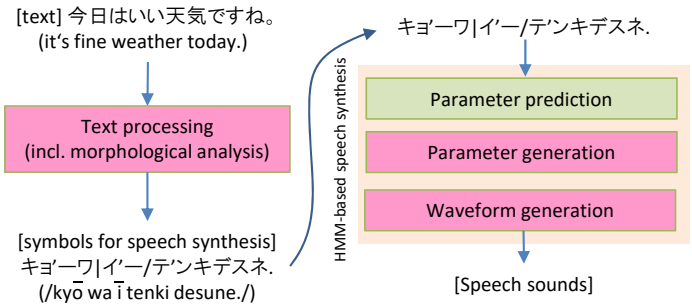


(back side) (4-layer circuit board)



There is no device for surface mount as a module

Processing for Japanese TTS



Techniques used for MCUs

Text processing:
- Morphological dictionary compressed by level-order unary degree sequence (LOUDS) [1]
  - Small but slow.
  - Fast text analysis is not necessary for TTS systems in practice.

HMM-based speech synthesis:
- Parameter generation with fixed-point arithmetic [2]
- Sinusoidal synthesis performed on a subband coding system for waveform generation [3]

- Pipelined processing with parameter generation and waveform generation
  - Not only for short latency but also for reduction of the RAM size

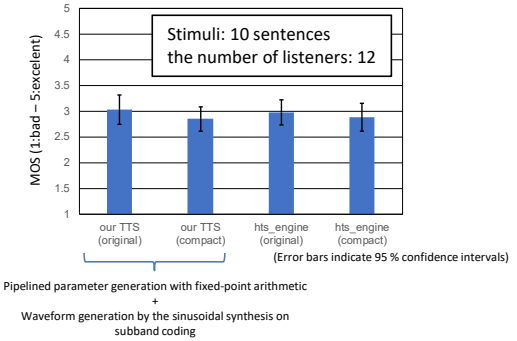
Configuration of our HMM-based synthesizer

Sampling frequency	32 kHz
Model parameters (output of HMMs)	31-order melcepstrum + Δ+ Δ^2, log F0 (with voiced/unvoiced information), state duration.
HMM structure	5-state left-to-right MSD-HMM
Label information	phoneme, length and position in the sentence in mora, distance from/to prosodic boundaries, accent information, etc.
Training data size	10.6-hour (by a female narrator)
Total number of states	Original: 1832 (duration), 7957 (mcep), 20995(log F0) Compact: 981 (duration), 5205 (mcep), 9694(log F0)
Model size	Original: 3.3MB*, Compact: 2.2MB

(*) Because the size of the morphological dictionary is 5.5MB, a compact model set was also built for the 8MB flash by change of the MDL parameter in the context clustering.

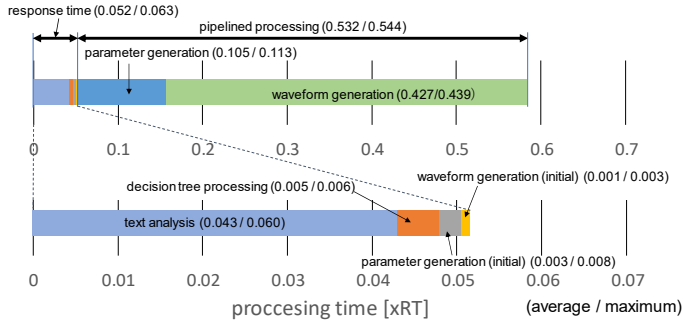
Subjective comparison

- No significant deterioration was not observed in comparison to hts_engine API 1.10 by NITECH.

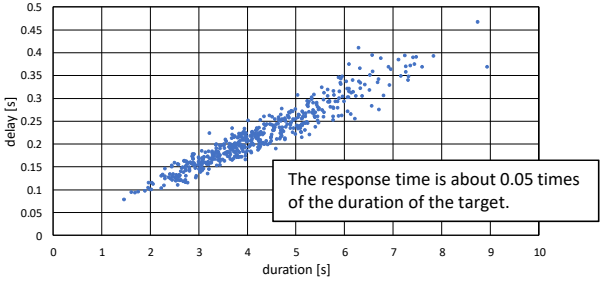


Processing speed

(for ATR503 phonemic-balanced sentences)



Response time of TTS processing
(without delay in D/A conversion)



Power consumption is <15mA when TTS is running

[1] O. Delpratt, N. Rahman and R. Raman, "Engineering the LOUDS Succinct Tree Representation," WEA 2006.
[2] N. Nishizawa and T.Kato, "Accurate parameter generation using fixed-point arithmetic for embedded HMM-based speech synthesizers," ICASSP 2011.
[3] N. Nishizawa and T.Kato, "Speech synthesis using a maximally decimated pseudo QMF bank for embedded devices," SSW8, 2013.