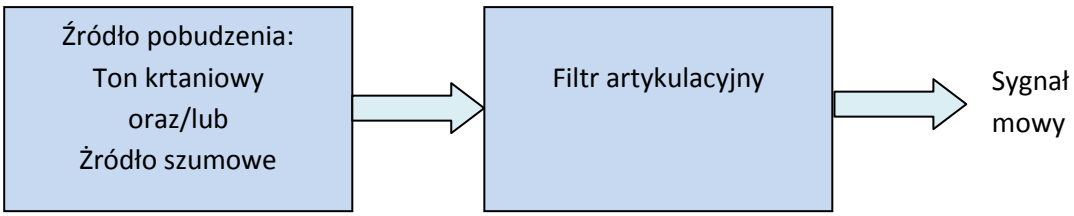
Jerzy Sawicki

Katedra Inżynierii Systemów, Sygnałów i Elektroniki
Wydział Elektryczny
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Pomiary i analiza parametru VOT w sygnale mowy

Parametr VOT (ang. Voice Onset Time) był początkowo [1] parametrem wprowadzonym do sterowania procesem syntezy mowy w układach formantowych. Z układami syntezy formantowej wiązano nadzieję na prawidłową produkcję mowy syntetycznej, gdyż generalnie biorąc układy te naśladują rzeczywiste procesy zachodzące podczas wytwarzania mowy przez człowieka. Układy syntezy formantowej bazują na dobrze znanym układzie typu źródło-filtr (rys. 1).

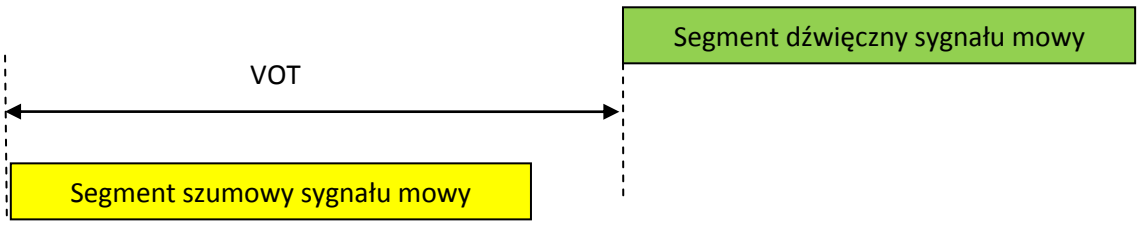


Rys. 1. Schemat blokowy procesu wytwarzania mowy typu źródło-filtr.

Badania i analizy rzeczywistego sygnału mowy wykazały poprawność i użyteczność takiego uproszczonego sposobu przedstawiania procesu wytwarzania, który zresztą mógł być i był uzupełniany o dodatkowe elementy zarówno w zakresie kształtowania sygnału, jak i bardziej złożonego procesu artykulacji mowy [2, 3]. Sprawdzenie poprawności modelu odbywało się także poprzez modelowanie i komputerowe symulacje zjawisk fizycznych zachodzących podczas wytwarzania mowy [4, 5]. Pomimo użyteczności praktycznej w wielu badaniach i analizach oraz swoich możliwości aplikacyjnych w technice, model ten dawał stosunkowo słabe rezultaty w układach elektronicznej syntezy mowy typu formantowego, a więc odtwarzających proces artykulacji mowy w sposób zbliżony do tego, co dzieje się w naturalnym procesie mówienia. Parametr VOT był jednym z ważnych parametrów syntezy formantowej, gdyż w sposób ilościowy decydował o sekwencji i odstępach czasowych włączania źródeł akustycznych: krtaniowego i szumowego. Pomimo wielu starań i prac

badawczych synteza mowy w oparciu o model formantowy nie dawała zadawalających rezultatów. Mowa syntezowana była daleka od naturalnej i męcząca przy konieczności jej słuchowej analizy. Synteza mowy z czasem została znacznie lepiej realizowana w układach konkatenacyjnych, w oparciu o naturalnie pozyskany materiał dźwiękowy. Przykładem takiego bardzo wydajnego syntezatora mowy jest platforma Iwona [6]. Tym niemniej parametr VOT zdefiniowany początkowo do celów syntezy mowy został uznany za użyteczny w środowisku fonetyków akustycznych, a więc badaczy zajmujących się analizą sygnału mowy w oparciu o ich parametry akustyczne (zazwyczaj czasowo-częstotliwościowe) do analizy mowy a w szczególności wykrywania jej zaburzeń. Takim właśnie celom poświęcone są badania prowadzone w Katedrze Systemów, Sygnałów i Elektroniki ZUT w Szczecinie we współpracy z dr. Lilianą Konopską z Uniwersytetu Szczecińskiego.

Parametr VOT jest wyrażonym w jednostkach czasu odstępem pomiędzy początkiem realizacji tzw. segmentu dźwięcznego mowy, a więc segmentu w którym występują regularne drgania fałdów głosowych krtani (objawiające się w przebiegu czasowym sygnału mowy zwiększonym poziomem sygnału, okresowością związaną z wysokością głosu, a w dziedzinie częstotliwościowej wyraźną strukturą harmoniczną wynikającą z okresowości drgań). Parametr VOT można zatem wyrazić schematycznie, jako odcinek czasu przedstawiony na wykresach czasowych sygnału mowy(rys. 2)



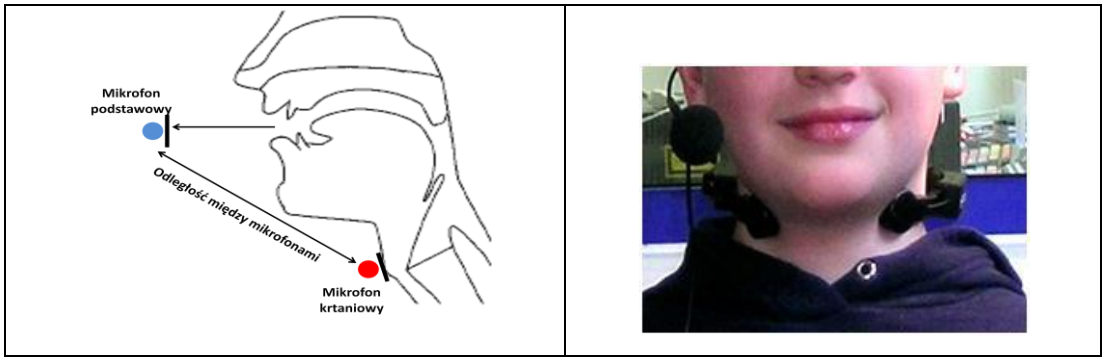
Rys. 2 Diagram czasowy przedstawiający definicję parametru VOT

Podczas wytwarzania sygnału mowy pojawiają się różne sekwencje następstw pomiędzy dwoma początkami dwóch podstawowych rodzajów sygnałów: dźwięcznego (krtaniowego) i szumowego (wybuchowego, plozyjnego). Możliwe i poprawne w konkretnych wypowiedziach są zarówno dodatnie, ujemne, jak i bliskie zera wartości parametru VOT.

Cykl badań parametru VOT u dzieci z zaburzeniami mowy przeprowadzono w Katedrze KISSE w 2012 roku. Sesje nagraniowe przeprowadzano w studyjnym pomieszczeniu Laboratorium Akustyki i Technologii Nagrań Dźwiękowych przy Katedrze Inżynierii Systemów, Sygnałów i Elektroniki Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie. W każdej sesji brały udział dwie osoby: inżynier dźwięku odpowiedzialny za rejestrację sygnału mowy i nadzorujący nagrania od strony technicznej, oraz osoba badająca przybywająca z dzieckiem w pomieszczeniu studyjnym. Wszyscy uczestnicy nagrań zostali poinformowani o ich celu i sposobie przeprowadzenia, mieli możliwość zapoznania się z osobami realizującymi sesje nagraniowe, pomieszczeniem studyjnym i sprzętem nagraniowym. Sesje nagraniowe przeprowadzono podczas jednorazowych spotkań, a czas trwania sesji nie przekraczał 20 minut. Do rejestracji sygnału mowy zastosowano dwa przetworniki: mikrofon podstawowy AKG C-555L z zausznymi zaczepami (nagłowny) oraz dodatkowy, podwójny mikrofon kontaktowy (krtaniowy, ang. *throat microph*one) LGF-24 firmy Navcomm, który umieszczano na szyi mówcy na wysokości krtani (powyżej chrząstki tarczowatej). Sygnał mowy z obu przetworników zapisywano równolegle na dwóch kanałach z częstotliwością próbkowania 44,1 kHz i w rozdzielczości 16 bitów na próbkę w rejestratorze cyfrowym Fostex FR-2 LE. Po zakończeniu nagrań cyfrowe pliki dźwiękowe w formacie .wav przenoszono na twardy dysk stacjonarnego komputera klasy PC.

Zastosowanie w rejestracji sygnału mowy pierwszego przetwornika – mikrofonu nagłownego zapewniło stałą odległość membrany od ust osoby mówiącej (w przybliżeniu 5 cm), co w przypadku mówców dziecięcych ma znaczenie, albowiem naturalna swoboda dziecięcych zachowań mogła w trakcie nagrań powodować niespodziewane i znaczne zmiany wartości poziomu dźwięku. Zastosowanie drugiego przetwornika podyktowane było tym, że mikrofon kontaktowy rejestruje drgania quasi-periodyczne nawet przy niskim poziomie amplitudy drgań krtaniowych, a przy wysokim poziomie zaszumienia sygnału zarejestrowanego przez mikrofon podstawowy dostarcza pełnych danych o przebiegu czasowym pobudzenia krtaniowego, co w prowadzonych badaniach miało szczególne znaczenie. Ze względu na wzajemne usytuowanie przetworników sygnał elektryczny w podstawowym mikrofonie był nieznacznie opóźniony w stosunku do sygnału z krtaniowego mikrofonu (rys. 3). Czasowe przesunięcie sygnałów w każdym pliku dźwiękowym dokładnie analizowano w programie do edycji, analizy i syntezy mowy *Praat*. Dla każdego pliku

dźwiękowego na podstawie 3-4 próbek pomiarowych quasi-periodycznego przebiegu sygnału mowy badanej osoby ustalano średnie opóźnienie czasowe, a następnie synchronizowano sygnały z obu kanałów i zapisywano w formacie .wav. Wszystkie uzyskane w badaniach pomiary wyznaczono podczas ręcznej segmentacji sygnału mowy w wersji 5.2 programu *Praat*.



Rys. 3. Sposób i miejsce umieszczenia przetworników elektroakustycznych

Pozyskany materiał dźwiękowy jest bardzo obszerny a jego opracowanie bardzo czasochłonne. Szczegółowe opracowanie uzyskanych wyników zostanie przedstawione w przygotowywanej publikacji, której współautorami będą dr Liliana Konopska (Uniwersytet Szczeciński w Szczecinie) oraz dr inż. Jerzy Sawicki (Zachodniopomorski Uniwersytet Technologiczny w Szczecinie).

Literatura

1. Lisker L., Abramson A. *A cross-language study of voicing in initial stops: Acoustical measurements*. Word, 20, s.384-422. (1964).
2. Jassem W. *Podstawy fonetyki akustycznej*. Wyd. PWN, Warszawa. (1973).
3. Jassem W. *Mowa a nauka o łączności*. Wyd. PWN, Warszawa. (1974).
4. Gubrynowicz R. *Komputerowe modelowanie artykulacji głosek języka polskiego*. Praca habilitacyjna. Instytut Podstawowych Problemów Techniki Polskiej Akademii Nauk, Warszawa. (2000).
5. Sawicki J. Sygnał mowy i możliwości jego korekcji metodami czasowo częstotliwościowymi/ Praca doktorska. Politechnika Szczecińska. Szczecin 1990.
6. Iwona. Text-to-speech. Syntezator mowy. www.iwona.com.pl