



KOMPRESJA A ZROZUMIAŁOŚĆ SYGNAŁU MOWY

Compression versus speech intelligibility

Hanna Wojciechowska

Institut Akustyki, Uniwersytet Im. Adama Mickiewicza,
ul. Umultowska 85, 61-114 Poznań,
hania@spl.ia.amu.edu.pl

STRESZCZENIE

Synteza sinusoidalna, w skrócie zwana SWS (SineWave Synthesis), poprzez odwzorowanie zmian dynamiki i położenia wybranych formantów w skali częstotliwości w czasie trwania wypowiedzi pozwala na znaczną redukcję danych zawartych w sygnale mowy (Remez, Rubin, Pisoni, & Carrell, 1981). Synteza ta odrzuca szczegółowe informacje akustyczne zawarte w sygnale. Jednocześnie, pomimo znacznej redukcji danych (współczynnik kompresji wynosi nawet 195:1), wiadomość zawarta w sygnale zostaje w dużym stopniu zachowana. W pracy wykorzystano po raz pierwszy syntezę SWS do analizy mowy polskiej w celu ustalenia zależności pomiędzy stopniem kompresji a zrozumiałością mowy. Najwyższy procent poprawnie zrozumianych słów uzyskano dla słów pojedynczych (ok. 85%). Dla zdań wypowiedzianych przez różnych mówców, przy maksymalnej kompresji, uzyskany wynik był o 20% gorszy. Podobnie w przypadku logatomów, procent poprawnie zrozumianych sylab był najwyższy dla logatomów prezentowanych pojedynczo. Przy maksymalnej kompresji nie przekroczył jednak 50%. Pokazano również, że nowa metoda SWS opracowana na UAM dała 2-krotnie lepsze rezultaty niż metoda opracowana w Instytucie Yale.

1. WPROWADZENIE

Z pojęciem sygnału akustycznego ściśle związane jest pojęcie informacji. Informacje zawarte w sygnale akustycznym dotyczą głównie własności samego źródła, własności fizycznych ośrodka i geometrii otoczenia w którym propaguje się fala akustyczna. Pomijając informacje osobnicze i emocjonalne, oraz lokalizacyjne mówcy względem odbiorcy, sygnał mowy niesie w sobie dodatkową informację jaką jest wiadomość.

Odwołując się do teorii widzenia D. Marra [3] i koncepcji opracowanej przez A. Klawitera [1], informację zawartą w sygnale akustycznym można podzielić na kilka poziomów. W przypadku sygnału mowy na poziomie najniższym zawarta jest informacja związana z treścią, czyli inaczej wiadomość przesłana przez mówcę do odbiorcy. Z punktu widzenia przekazu wiadomości poziomy wyższe są zbędne. Słuszność tego podejścia potwierdzają wyniki badań przeprowadzonych w roku 1982 przez R. Remeza i P. Rubina [4]. Poddali oni sygnał mowy syntezie sinusoidalnej, SWS (z ang. SineWave Synthesis). W syntezie tej wyeksponowana grupa składowych widma dźwięku, np. formanty w sygnale

mowy, jest modelowana za pomocą kilku tonów oddających zmiany dynamiki i lokalizacji w skali częstotliwości w czasie trwania dźwięku. Odtworzone w ten sposób sygnały mowy tracą swoją naturalność; zostają w dużym stopniu pozbawione informacji prozodycznych i lokalizacyjnych. Pozostaje pytanie, w jakim stopniu zostaje zachowana w nich wiadomość.

2. CEL PRACY

Głównym celem badań było określenie związku pomiędzy stopniem kompresji sygnału mowy a procentem poprawnie zrozumianej wiadomości wypowiedzianej w języku polskim. Przeprowadzone analizy miały na celu określenie w jakim stopniu poprawność odbioru wiadomości jest związana z czasem prezentacji sygnału, jego strukturą gramatyczną i znaczeniową. Przeprowadzono 5 eksperymentów, w których słuchaczom prezentowano materiał testowy w postaci zdań, wyrażeń utworzonych z 3 słów, słów pojedynczych, wyrażeń utworzonych z 3 logatomów i pojedynczych logatomów. Celem ostatniego eksperymentu było porównanie jakości metody SWS zastosowanej w tej pracy z oryginalną metodą opracowaną w Instytucie Yale.

3. METODA BADAŃ

W eksperymentach zastosowano nową, opracowaną na UAM metodę wyszukiwania i odwzorowywania formantów, uwzględniającą rozkład energii w sygnałach mowy polskiej. Ze względu na dużą liczbę dźwięków, których energia zawiera się w wysokim zakresie częstotliwości, zakres wyszukiwania formantów objął pasmo od 100 Hz do 8 kHz. W metodzie odwzorowywano formanty o najwyższej amplitudzie. Efektywny odstęp między odwzorowywanymi zmiennymi w czasie formantami wynosił 20 ms. Wszystkie sygnały był prezentowane słuchaczom na poziomie 65dB SPL.

3.1 Materiał testowy

W eksperymentach zastosowano zróżnicowany pod względem gramatycznym jak i pod względem zawartej informacji materiał testowy. Sygnałami testowymi były zarówno zdania (wybrane z bazy nagrań sygnałów mowy- CORPORY [2]) wyrazy jak i logatomy. Wyrazy pochodziły z bazy nagrań słów zrównoważonych pod względem liczby formantów o wysokiej i niskiej częstotliwości opracowanej i udostępnionej przez prof. W. Jassemę. Listy logatomowe udostępnił dr inż. S. Brahmański.

3.2 Procedura

W każdym eksperymencie słuchacze w specjalnym oknie dialogowym zapisywali to co usłyszeli. Na podstawie uzyskanych w ten sposób wyników określono procent zrozumiałości słów (w przypadku wyrażeń słownych) i procent zrozumiałości sylab (w przypadku wyrażeń utworzonych z logatomów).

4. PRZEBIEG EKSPERYMENTU

W pierwszym eksperymencie słuchacze oceniali zrozumiałość 108 dwusekundowych zdań podzielonych na 4 listy o różnym stopniu kompresji. Słuchacze zostali podzieleni na 2 grupy. Pierwszej grupie prezentowane były zdania wygenerowane przez różnych mówców, a drugiej grupie prezentowane były zdania wypowiedziane przez jednego mówcę. Słuchacze najpierw słuchali sygnałów zsyntetyzowanych za pomocą 9 tonów, potem kolejno 6, 4 i 3 tonów. Między kolejnymi odsłuchami zachowany był odstęp czasowy - minimalnie wynosił on 30 min. Odsłuchy zostały przeprowadzone w takiej kolejności, ponieważ sygnały syntetyczne na początku wydają się bardzo nienaturalne a przez to trudno „wychycić” wiadomość w nich zawartą. W chwili gdy układ słuchowy przyjmie sygnały syntetyczne jako sygnały mowy słuchacz znacznie lepiej zaczyna percypować informację. Ponieważ proces „uczenia” jest niezwykle szybki sygnały zsyntetyzowane z 9 tonów przeznaczone były na „trening” i „zapoznanie” słuchaczy z sygnałami.

W drugim eksperymencie słuchacze oceniali zrozumiałość 27-wypowiedzi, każda złożona z trzech rzeczowników, skompresowanych na poziomie 3 tonów. Długość każdej wypowiedzi wynosiła ok. 2 sekund. Wszystkie słowa zostały wygenerowane przez jednego mówcę.

W eksperymencie trzecim słuchacze słuchali tych samych słów ale prezentowanych pojedynczo. Stopień kompresji został zachowany na poziomie trzech tonów. Słowa podzielono na 2 listy, po 40 słów każda.

W czwartym eksperymencie słuchaczom zaprezentowano 108 dwusekundowych wyrażen, każde składające się z trzech logatomów. Wyrażenia te zostały podzielone na 4 listy o różnym stopniu kompresji. Stopień kompresji i kolejność odsłuchów były takie jak w przypadku zdań. Tym razem nie chodziło jednak o „trening” lecz raczej uwzględniono poziom trudności jaki sprawiają odsłuchy logatomów, nawet w przypadku gdy są odtwarzane jako sygnały nieskompresowane.

W piątym eksperymencie słuchacze słuchali pojedynczych logatomów zsyntetyzowanych na poziomie 3 tonów. Lista odsłuchowa składała się z 50 logatomów.

W szóstym eksperymencie słuchaczom przedstawiono zdania z pierwszego eksperymentu, ale zsyntetyzowane zgodnie z metodą opracowaną w Instytucie Yale. Parametry w programie zostały tak dobrane, by wygenerowane zdania miały ten sam poziom kompresji co sygnały wygenerowane metodą opracowaną w Instytucie Akustyki UAM.

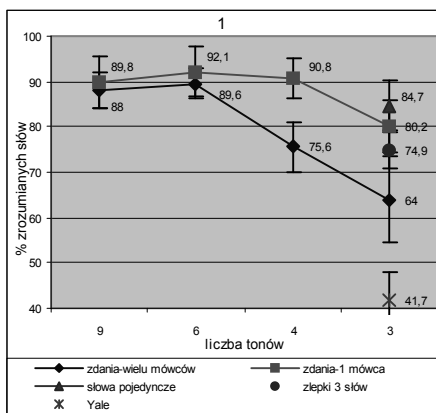
5. REZULTATY I DYSKUSJA WYNIKÓW

Na rysunku 1 przedstawiono rezultaty eksperymentów 1-3 i 6, jako procent poprawnie zrozumianych słów, uśredniony po 40 słuchaczach, w funkcji liczby tonów odwzorowujących sygnał. Największy procent zrozumiałości w przypadku słów uzyskano nie przy 9 tonach, lecz przy 6 i to niezależnie od liczby mówców generujących sygnał. Można to wytłumaczyć biorąc pod uwagę fakt, że prezentowane sygnały odznaczały się dużą nienaturalnością i słuchacze zwracali uwagę nie tyle na wiadomość zawartą w sygnale co na jego brzmieniu. Dla każdego poziomu kompresji uzyskane wyniki były tym lepsze, im krótsza była wiadomość i im mniej dodatkowej informacji niósł sygnał. Najlepsze wyniki uzyskano dla pojedynczych słów, wygenerowanych przez jednego mówcę. Gdy wypowiedź stała się dłuższa, zrozumiałość wiadomości automatycznie zmalała. Łatwiej było zrozumieć niepowiązane ze sobą w żaden sposób oderwane słowa wypowiedziane przez jednego mówcę, niż tak samo długo trwające logiczne zdania ale wygenerowane przez różnych mówców. Zatem dodatkowa informacja, jaką jest identyfikacja mówcy, znacznie pogarsza odbiór wiadomości. Układ słuchowy najpierw interpretuje informację z wyższego poziomu,

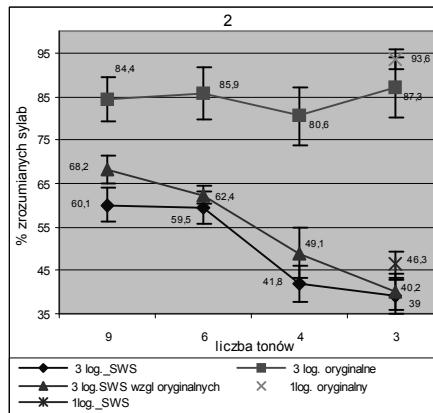
czyli identyfikuje mowę, a dopiero potem przechodzi na poziom niższy, czyli „odbiera” wiadomość. Analogiczna sytuacja ma miejsce w percepcji wzrokowej - w pierwszej kolejności przyciąga nas obraz graficzny (wyższy poziom informacji) a dopiero potem treść (najniższy poziom informacji). Z drugiej strony wydawać by się jednak mogło, że zdania ciągle powinny być lepiej zrozumiane od oderwanych słów, a tak nie jest. Tutaj znaczenie mogła mieć szybkość wypowiedzianych słów i fakt, że prezentowane zdania były bardzo nisko redundantne. Słuchacze często popełniali błąd szukając sensu logicznego i „na siłę” zmieniali znaczenie słowa, by ten sens uzyskać.

Na rysunku 2 przedstawiono rezultaty eksperymentów 4-5, jako procent poprawnie zrozumianych słów, uśredniony po 20 słuchaczach, w funkcji liczby tonów odwzorowujących sygnał. W przypadku logatomów zrozumiałość wyniosła poniżej 50%. Należy jednak zaznaczyć, że zrozumiałość logatomów nieskompresowanych wahała się w granicach 85-90 %, podczas gdy dla oryginalnych słów i zdań wyniosła 100%. Podobnie jak w eksperymencie słownym, najlepszą zrozumiałość osiągnięto dla logatomów prezentowanych pojedynczo.

Metoda ta może znaleźć zastosowanie w dalszych poszukiwaniach nowych metod kompresji sygnału, zwłaszcza w sytuacjach gdzie ważna jest jedynie wiadomość zawarta w zdaniu a nie informacje związane z osobą mówcy. Innymi słowy, uzyskaliśmy odpowiednik informacji przesyłanych alfabetem Morsa XXI wieku.



Rys. 1. Procent zrozumianych słów w zależności od poziomu kompresji.



Rys. 2. Procent zrozumianych sylab w zależności od poziomu kompresji.

LITERATURA

1. A. Klawiter, O słyszeniu przedmiotów , Umysł a Rzeczywistość, Poznańskie Studia z Filozofii Humanistyki, 5 (18), 1999, 327-339.
2. Baza nagrań sygnałów mowy CORPORA, wykonana w ramach KBN 8T11C 023 8, kierownik projektu S. Grocholewski
3. D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, W.H. Freeman, San Francisco, 1982.
4. R.E. Remez, P.E. Rubin, D.B. Pisoni, T.D. & Carrell, Speech perception without traditional speech cues. Science, 1981, 212, 947-950.