XXVIII Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE 2018

Oddział Gdański PTETiS

doi: 10.32016/1.60.09

POPRAWA OBIEKTYWNYCH WSKAŹNIKÓW JAKOŚCI MOWY W WARUNKACH HAŁASU

Krzysztof KĄKOL[1], **Bożena KOSTEK**[2]

1. Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
   e-mail: info@creasoft.pl
2. Laboratorium Akustyki Fonicznej, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
   e-mail: bokostek@audioacoustics.org

Streszczenie: Celem pracy jest modyfikacja sygnału mowy, aby uzyskać zwiększenie poprawy obiektywnych wskaźników jakości mowy po zmiksowaniu sygnału użytecznego z szumem bądź z sygnałem zakłócającym. Wykonane modyfikacje sygnału bazują na cechach mowy lombardzkiej, a w szczególności na efekcie podniesienia częstotliwości podstawowej F0. Sesja nagraniowa obejmowała zestawy słów i zdań w języku polskim, nagrane w warunkach ciszy, jak również w obecności sygnałów zakłócających, tj. szumu różowego oraz tzw. gwaru (ang. *babble speech*), określanego też jako efekt „cocktail-party". W ramach badań przetwarzano próbki mowy głosów męskich. W pracy wykazano, że podniesienie częstotliwości podstawowej skutkuje zwiększonymi wartościami wskaźnika jakości mowy, mierzonymi przy użyciu standardu PESQ (*Perceptual Evaluation of Speech Quality*).

Słowa kluczowe: efekt Lombarda; wskaźnik oceny jakości sygnału mowy PESQ (*Perceptual Evaluation of Speech Quality*); parametry sygnału mowy.

1. WPROWADZENIE

Mowa lombardzka jest efektem odkrytym w roku 1909 przez Etienne Lombarda, francuskiego otolaryngologa [1]. Efekt ten występuje w sytuacji, kiedy mówca w sposób nieuświadomiony zmienia pewne cechy akustyczne własnej mowy w hałasie. Lombard zaobserwował, że np. mówcy w obecności tłumu wypowiadają się w nieco inny sposób niż wtedy, gdy mają okazję mówić w sytuacji bardziej kameralnej.

W licznych badaniach nad mową lombardzką [2][3][4][5][6] zidentyfikowano wiele cech charakterystycznych dla tego typu wypowiedzi [3][7][8][9], m.in. podniesienie częstotliwości podstawowej czy przesunięcie energii z pasm o niższych częstotliwościach do częstotliwości średnich i wyższych. W niniejszej pracy skupiono się przede wszystkim na zwiększeniu wartości częstotliwości podstawowej.

Prace nad mową lombardzką często koncentrują się na subiektywnych badaniach zrozumiałości mowy. Istnieją jednak wskaźniki obiektywne takie jak PESQ (*Perceptual Evaluation of Speech Quality*) czy P.563, które znajdują zastosowanie w badaniach jakości kanałów telekomunikacyjnych [10][11][12][13].

Badania przeprowadzone w ramach niniejszej pracy obejmowały kilka etapów: (1) nagrania próbek dźwiękowych (zdania) bez i w obecności szumu różowego i zakłóceń (*babble speech*), tj. w pierwszej kolejności nagrywano sygnał referencyjny, a następnie nagrywano te same zdania w obecności dodatkowych zakłóceń wymuszając efekt Lombarda w nagraniach mowy; (2) w celu pomiaru współczynników PESQ MOS nagrania te zostały następnie zmiksowane z szumem różowym z różną wartością stosunku sygnału do szumu (SNR); (3) w pomiarze wskaźnika PESQ wykorzystano pliki źródłowe (pliki referencyjne) przetworzone poprzez podniesienie wartości F0 i poziomu natężenia dźwięku oraz te same pliki zmiksowane z szumem różowym i sygnałem zakłócającym *babble speech*.

W niniejszej pracy przetwarzano sygnał mowy nagrany w warunkach ciszy w taki sposób, aby jego cechy (głównie częstotliwość podstawowa) odpowiadały cechom mowy lombardzkiej. Wykazano też, że takie przetwarzanie poprawia obiektywne wskaźniki jakości mowy.

2. CECHY MOWY LOMBARDZKIEJ

Cechy mowy lombardzkiej są różnorodne i obejmują następujące zjawiska [2][3][7][8][9][14]:

- zwiększenie poziomu natężenia dźwięku,
- podniesienie częstotliwości podstawowej sygnału,
- przesunięcie energii z pasm o niższej częstotliwościach do pasm o częstotliwościach wyższych,
- wzrost wartości formantów, głównie F1 i F2,
- zwiększenie długości trwania samogłosek,
- zwiększenie nachylenia widma (tzw. *spectral tilt*).

Większość tych cech można łatwo wyznaczyć, jednak obserwacja zmian tych cech w kontekście mowy lombardzkiej nie jest już taka prosta. Pomiar chwilowej wartości częstotliwości formantów nie jest w tym przypadku miarodajny, gdyż może wskazywać na przykład na chwilową zmianę związaną np. z emocjami zawartymi w wypowiedzi. W przypadku detekcji mowy lombardzkiej łatwiej więc stosować miary długookresowe, np. medianę wartości częstotliwości podstawowej w dłuższym czasie.

3. OBIEKTYWNE WSKAŹNIKI JAKOŚCI MOWY

Z punktu widzenia jakości mowy najlepszym miernikiem rzeczywistej jakości jest subiektywny pomiar zrozumiałości. ITU definiuje standardy pomiarów subiektywnych – z wykorzystaniem paneli słuchaczy. Najważniejsze normy zostały zestawione w formie standardu P.800/P.830 [12][15]. Rezultaty tego typu pomiarów są przedstawiane jako MOS (*Mean Opinion Score*) [15]. Pomiar taki powinien być wykonany w formie testów odsłuchowych w grupie słuchaczy. Pomimo zestandaryzowania tego typu testów (wymagania dotyczące akustyki wnętrza

odsłuchowego, dopuszczalnej wartości zakłóceń, systemu odsłuchowego, sposobu prowadzenia testów, wiarygodności testerów, itp., zawarte w normach, m.in.: ITU-R BS.1116 oraz ITU-R BS.1284 [16][17]), są one obarczone z reguły błędami wynikającymi z faktu, że każdy słuchacz może mieć inne doświadczenia słuchowe. Należy jednak pamiętać, że zrozumiałość mowy jest z definicji wskaźnikiem subiektywnym, dlatego tego typu badania prowadzone są jako ostateczna weryfikacja wskaźników.

Obiektywnie można mierzyć jedynie jakość sygnału mowy [10]. Wskaźniki obiektywne dotyczące sygnału mowy (np. wyrazistość) wykorzystywane są najczęściej w pomiarach jakości kanałów telekomunikacyjnych. Pomiar jakości mowy jest istotny przede wszystkim ze względu na konieczność zapewnienia właściwej jakości świadczonych usług. Standardy pomiaru stosuje się głównie do kanałów telekomunikacyjnych, ponieważ zapewnienie właściwej jakości kanału transmisyjnego jest bardzo istotne z punktu widzenia optymalizacji kosztu jego wykorzystania.

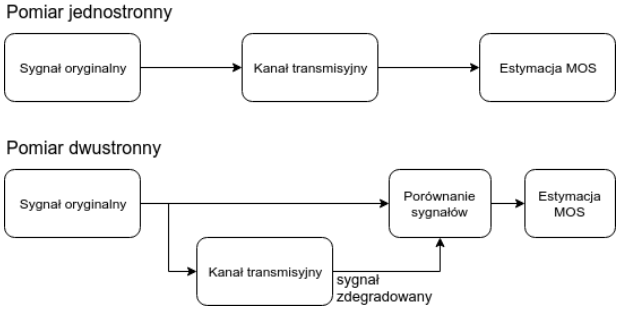
Istnieje wiele czynników wpływających na jakość sygnału mowy, są to m.in.:

- wąskie pasmo transmisji lub kodowanie przy użyciu niskiej szybkości transmisji,
- algorytmy kompresji i kodowania,
- szum tła,
- opóźnienie pakietów w transmisji cyfrowej,
- jakość urządzeń transmisyjnych (np. telefonów komórkowych).

Jakość mowy w kanałach telekomunikacyjnych można mierzyć na dwa sposoby:

- pomiar dwustronny – wykonywany przez porównanie sygnału przed kanałem i za kanałem (sygnał referencyjny i testowany),
- pomiar jednostronny – szacowanie jakości mowy w oparciu o czynniki perceptualne, bez znajomości sygnału referencyjnego.

Porównanie obu metod pomiarów zamieszczono na rysunku 1.



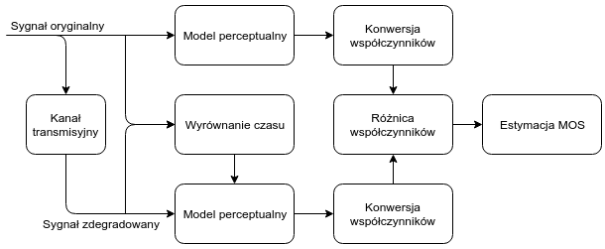
Rys. 1. Porównanie metod pomiaru jednostronnego i dwustronnego [11]

3.1. Pomiary dwustronne

Pierwszym standardem pomiaru jakości sygnału mowy był P.861, znany jako PSQM. Niestety standard ten nie uwzględniał wielu czynników obecnych we współczesnych cyfrowych kanałach transmisyjnych, np. utraty pakietów w transmisji VoIP, szumu tła czy też zmiennego opóźnienia. Wszystkie te czynniki zostały uwzględnione w standardzie P.862 (PESQ) [13]. W pomiarach porównawczych testów obiektywnych i subiektywnych współczynnik korelacji testu PSQM z pomiarami subiektywnymi osiągał wartość 0.26, zaś w przypadku analizy PESQ – 0.93.

Schemat działania algorytmu PESQ przedstawiony jest na rysunku 2. Model ten obejmuje następujące etapy:

- wyrównanie poziomu sygnału; w przypadku porównywania sygnału oryginalnego i zdegradowanego, należy wyrównać oba do tego samego poziomu mocy,
- filtrowanie wejściowe; w PESQ modelowane jest filtrowanie sygnału, które odbywa się w urządzeniach telefonicznych i w sieci telekomunikacyjnej,



Rys. 2. Schemat działania algorytmu PESQ [11]

- wyrównanie czasowe; systemy transmisyjne (np. VoIP) wprowadzają do sygnału znaczące opóźnienie, które musi być skompensowane przed pomiarem,
- transformacja słuchowa (perceptualna); sygnał referencyjny i zdegradowany przetwarzany jest przez system transformacji słuchowej, aby zasymulować cechy ludzkiego słyszenia - przykładowo system ten usuwa te części sygnału, które nie są słyszalne przez słuchacza,
- obliczanie zakłóceń.

Obliczone parametry zakłóceń są konwertowane na wynik PESQ w zakresie od -1 do 4.5. Standardowo wyniki te przeliczane są na skalę MOS-LQO (*Listening Quality Objective*) [18], czyli wartości od 1 do 5 (gdzie 1 – oznacza złą jakość mowy, a 5 - bardzo dobrą jakość).

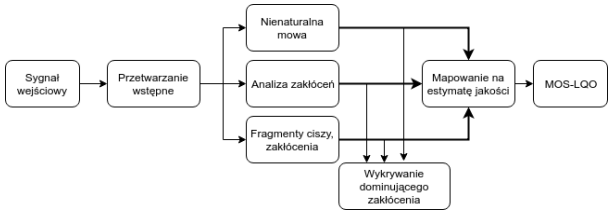
3.2. Pomiary jednostronne

W pomiarach jednostronnych szacuje się wartość MOS wyłącznie na podstawie sygnału z zakłóceniami. W przypadku standardu P.563 należy zasymulować wykorzystanie rzeczywistego eksperta, odsłuchującego rozmowę na urządzeniu testowym. Tym urządzeniem może być dowolny odbiornik, np. telefon komórkowy. Ponieważ w tym przypadku nie porównuje się zdegradowanego sygnału z sygnałem oryginalnym, wskaźnik jakości mowy zależy od urządzenia odsłuchowego. Jest ono więc istotnym elementem standardu P.563 [11].

Każdy sygnał poddany pomiarowi MOS przy użyciu P.563 musi zostać wstępnie przetworzony, poprzez wykorzystanie modelu urządzenia odsłuchowego. Następnie wykorzystany jest detektor mowy (VAD - *Voice Activity Detector*), aby oznaczyć fragmenty sygnału związane z mową. W kolejnym etapie sygnał mowy jest poddawany szeregowi analiz i przyporządkowywany do określonej klasy zakłóceń.

Parametryzacja sygnału w P.563 może być podzielona na trzy podstawowe bloki funkcyjne (schemat przedstawiono na rysunku 3) [11]:

- analiza traktu głosowego i nienaturalności mowy; można w tym przypadku wyróżnić wskazanie nienaturalności osobno dla głosów żeńskich i męskich oraz tzw. efekt "robota",
- analizę dodatkowego szumu; w tym przypadku ważna jest detekcja statycznego szumu tła oraz szumu związanego z obwiednią sygnału,
- przerwania, wyciszenia i obcięcia czasowe.



Rys.3. Schemat działania algorytmu P.563 [11]

Sygnał testowy musi także spełniać określone w standardzie wymagania, aby istniała możliwość detekcji jakości mowy przy użyciu algorytmu P.563, m.in.:

- częstotliwość próbkowania musi być większa lub równa 8 kHz,
- rozdzielczość sygnału cyfrowego musi być 16-bitowa
- sygnał nie może być dłuższy niż 20 sekund, a mowa w sygnale nie może być krótsza niż 3 sekundy.

3.3. Metoda pomiaru PESQ MOS

Jak wspomniano wcześniej, PESQ (*Perceptual Evaluation of Speech Quality*) jest miarą jakości sygnału w kanale telekomunikacyjnym [13]. W niniejszej pracy wykorzystano implementację algorytmu PESQ dostępną na stronach ITU. Implementacja ta pozwala przeprowadzić pomiary PESQ poprzez porównanie dwóch sygnałów, z których jeden jest sygnałem oryginalnym (referencyjnym), a drugi sygnałem testowym - z zakłóceniami (w przypadku prowadzonych badań – sygnałem z domiksowanym szumem różowym lub sygnałem *babble speech*, imitującym gwar głosów ludzkich).

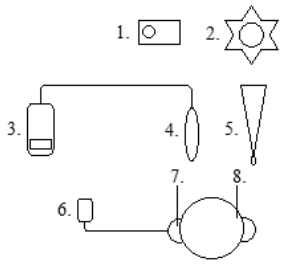
3.4. Metoda pomiaru P.563 MOS

Również w przypadku pomiaru P.563 w ramach eksperymentów wykorzystano implementację tego algorytmu dostępną na stronach ITU. Implementacja ta pozwala na wykonanie pomiarów z wykorzystaniem sygnału z zakłóceniami i wykonuje oszacowanie współczynnika jakości mowy.

4. METODOLOGIA NAGRAŃ

Nagrania próbek dźwiękowych dokonano w sali bez dodatkowych przegród akustycznych. Sprzęt nagraniowy i pomocniczy został rozmieszczony tak, jak pokazano na rysunku 4. Ze względu na większy zakres badań nagrania były audiowizualne, ale w ramach przedstawianych analiz skoncentrowano się wyłącznie na aspekcie dźwiękowym. Kamera oraz lampa oświetlająca mówcę zostały umieszczone ok. 2 m od mówcy. Ze względu na uwarunkowania audiowizualne (zapewnienie pełnej widoczności ust w obiektywie) mikrofon kierunkowy i miernik, skierowane w stronę ust nagrywanej osoby, umieszczono w odległości ok. metra. Mówcy nagrywani byli w pozycji siedzącej, ustawienia mikrofonu i kamery były korygowane w zależności od wzrostu danej osoby.

Przeprowadzone nagrania wykonane zostały bez obecności dodatkowych zakłóceń (sygnał referencyjny oraz w obecności szumu różowego i sygnału *babble speech* (dźwięk nagrany w restauracji, pobrany z Internetu). Szum różowy został wygenerowany przy pomocy aplikacji Noise Generator Soft. Przyjęto progi 70% oraz 90% wartości poziomu natężenia sygnału z odtwarzacza, czyli kolejno ok. 72,5 dB oraz 83,8 dB.



Rys.4. Schemat rozmieszczenia urządzeń: 1) kamera, 2) lampa, 3) rejestrator dźwięku, 4) mikrofon, 5) miernik, 6) odtwarzacz dźwięku, 7) słuchawki, 8) mówca

Nagrano 17 zdań różnego typu (z różną prozodią) oraz dodatkowo 10 osobnych słów, odpowiednio dla kobiet i mężczyzn. W niniejszej analizie posłużono się sygnałami zawierającymi zdania dla głosów męskich. Badania w niniejszej pracy zawierają więc wyniki analiz uzyskanych z eksperymentów wykonanych dla 17 różnych wypowiedzi oznajmujących, rozkazujących i pytających.

5. PORÓWNANIE JAKOŚCI MOWY W NAGRANIACH

Dla potrzeb niniejszej pracy wykorzystano w pierwszej kolejności nagrania wykonane w obecności zakłóceń. Ścieżka dźwiękowa nagrania nie zawiera jednak szumu – zgodnie z opisem przedstawionym w rozdziale 4. Nagrania te zostały następnie zmiksowane z szumem różowym z różną wartością stosunku sygnału do szumu (SNR). Dla kolejnych wersji badanych próbek SNR wynosił odpowiednio: -10 dB, -5 dB, 0 dB, 5 dB, 10 dB.

Wykorzystano zestaw nagrań głosu męskiego w języku polskim, które zostały wykonane w dwóch wersjach:

- bez obecności szumu,
- z szumem różowym o natężeniu 83,8 dB.

W obu przypadkach wykonano porównanie jakości mowy dla różnych wariantów poziomu zmiksowanego szumu. Wyniki przedstawiono w tabeli 1.

Widać wyraźnie, ze mowa lombardzka wpływa na poprawę obiektywnego wskaźnika jakości mowy (PESQ). Poprawa jest znacząca i zależy od wartości SNR. Dla dużych poziomów zakłóceń (SNR na poziomie -10 dB) MOS jest większy nawet o 62%, podczas gdy dla niskich poziomów (SNR na poziomie 10 dB) MOS jest większy o ok.16%. Średnia poprawa współczynnika MOS dla wykonanych porównań wynosi 37%.

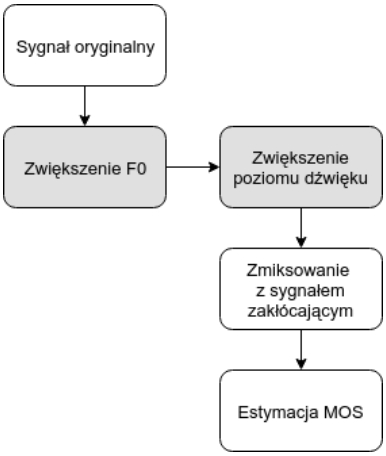
Tabela 1: Porównanie współczynników PESQ MOS dla wykonanych nagrań i różnego poziomu zmiksowanego szumu

Poziom SNR zmiksowanego zakłócenia [dB]	Nagranie w ciszy – MOS	Nagranie z szumem różowym 83 dB – MOS
-10	0,7427	1,2030
-5	0,7403	1,1850
0	0,9462	1,2322
5	1,3131	1,5515
10	1,6565	1,9196

6. MODYFIKACJE NAGRAŃ - WSKAŹNIKI JAKOŚCI MOWY

Biorąc pod uwagę czynniki decydujące o tym czy dane nagranie wypowiedzi zakwalifikować jako mowę lombardzką, czy też nie, w niniejszej pracy zdecydowano się skoncentrować na zmianie wysokości częstotliwości podstawowej (F0) oraz na zwiększeniu poziomu natężenia dźwięku.

Oba te czynniki zmieniane były w oparciu o wynik pomiaru różnicy natężenia i F0 w nagraniach głosu męskiego dla nagrania bez towarzyszącego szumu różowego oraz z tym szumem na poziomie 83,8 dB. Na rysunku 5 przedstawiono schemat wykonanych modyfikacji sygnałowych i przeprowadzane pomiary.



Rys.5. Schemat dokonywanych zmian w sygnale

6.1. Zwiększenie częstotliwości podstawowej

F0 zmienia się w trakcie pojedynczego nagrania, a także w danej wypowiedzi występują fragmenty, w których nie da się wyznaczyć F0 (np. w momentach pauz lub wystąpienia bezdźwięcznych fonemów), dlatego nie posłużono się pojedynczą wartością chwilowego F0. Wartością zmienianą była natomiast mediana F0 wyznaczona dla całego (kilkusekundowego) nagrania. Następnie przetwarzano nagranie w taki sposób, aby podnieść F0 o konkretną wartość. W praktyce uzyskana wartość była proporcjonalna do obliczonej mediany.

W ramach eksperymentów podnoszono wartość częstotliwości podstawowej od 10% do 60% ze skokiem 10%. Najlepsze efekty poprawy jakości mowy są zależne od rodzaju zakłóceń zmiksowanych z sygnałem wejściowym. Średnia różnica częstotliwości podstawowej w nagraniach bez i z mową lombardzką wyniosła 42.6%.

6.2. Zwiększenie poziomu natężenia dźwięku

Drugą charakterystyczną cechą nagrań mowy lombardzkiej, wykorzystywaną w badaniach, było zwiększenie poziomu natężenia dźwięku. Natura tego efektu jest dosyć oczywista – przebywając w hałasie, człowiek mówi głośniej. Średnia różnica w natężeniu dźwięku badanych nagrań wyniosła ok. 12% i o tę właśnie wartość podnoszono poziom dźwięku nagrania wejściowego, już po zwiększeniu F0.

W ten sposób przetworzony dźwięk (tzn. ze zwiększoną częstotliwością podstawową i zwiększonym poziomem) został zapisany jako plik referencyjny, służący do obliczenia współczynnika PESQ.

6.3. Zmiksowanie nagrania z szumem różowym i hałasem otoczenia

Po tej dwuetapowej obróbce nagranej wypowiedzi domiksowano do niej dwa rodzaje sztucznych sygnałów zakłócających:

- szum różowy,
- sygnał zakłócający *babble speech*, czyli hałas typowy dla miejsc, w których przebywa dużo ludzi.

Oba rodzaje zakłóceń zostały zmiksowane z sygnałem oryginalnym przy zachowaniu takich wartości SNR, jak przy porównaniu nagrań bez i z mową lombardzką, tj. -10, -5, 0, 5 oraz 10 dB.

6.4. Wyznaczenie obiektywnych współczynników jakości mowy

Dla przetworzonego w ten sposób sygnału obliczono obiektywne wskaźniki jakości mowy. Dla obu wskaźników przyjętych w niniejszym opracowaniu (tj. P.563 i PESQ) porównanie odbywa się w inny sposób ze względu na inną specyfikę zastosowanych algorytmów.

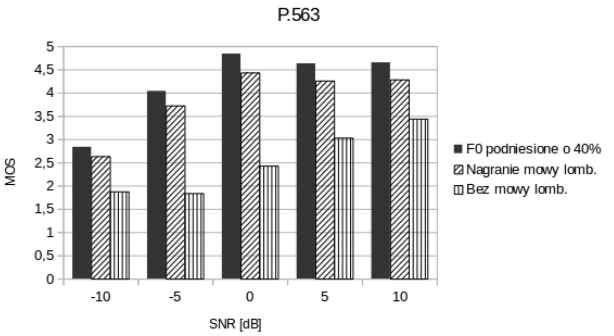
6.4.1. P.563

Dla P.563 potrzebna jest dodatkowa operacja przepróbkowania sygnału do niższej częstotliwości próbkowania, najlepiej do wartości 16 kHz. Przy zastosowaniu częstotliwości 48 kHz aplikacja do obliczania P.563 w niepoprawny sposób zmniejsza po raz kolejny częstotliwość do 8 kHz i zwraca błędne wartości współczynnika MOS.

Pomiar wartości MOS przy użyciu algorytmu P.563 jest jednostronny, tzn. nie jest wykorzystany sygnał referencyjny, a pomiar odbywa się poprzez uruchomienie aplikacji dla konkretnego, testowanego nagrania.

Pomiary P.563 zostały wykonane wyłącznie dla nagrań zmiksowanych z szumem różowym oraz przy podniesieniu częstotliwości podstawowej o 40%. Pomiary wskazują na dużą poprawę jakości sygnału, tym niemniej z uzyskanych danych wynika, że podniesienie F0 daje lepsze efekty niż wykorzystanie nagrań mowy lombardzkiej, co powinno zostać zweryfikowane w testach odsłuchowym, gdyż wydaje się wątpliwe. Dlatego też pomiar P.563 został odrzucony w toku wykonywanych eksperymentów.

Wyniki eksperymentów z P.563 zostały przedstawione na rysunku 6.



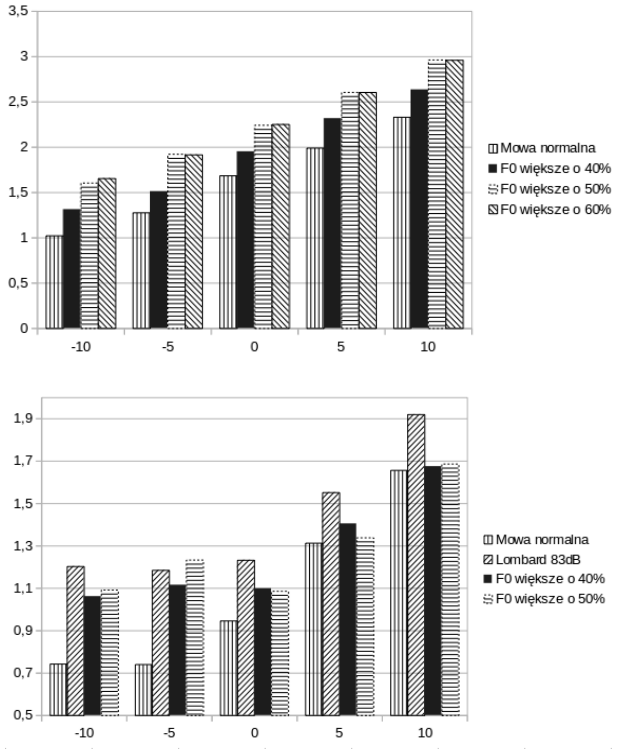
Rys.6. Poprawa wskaźnika P.563 MOS dla F0 podniesionego o 40% w zależności od SNR szumu towarzyszącego nagraniom

6.4.2. PESQ

Do pomiarów PESQ wykorzystano pliki źródłowe przetworzone poprzez podniesienie wartości F0 i poziomu natężenia dźwięku (pliki referencyjne) oraz te same pliki zmiksowane z sygnałem zakłócającym *babble speech* i szumem różowym (pliki badane). Pomiar PESQ odbywa się poprzez podanie na wejściu algorytmu pliku referencyjnego oraz „zdegradowanego" (w tym przypadku był to plik badany z domiksowanym szumem lub sygnałem zakłócającym).

Wyniki pomiarów wskazują na poprawę współczynnika MOS PESQ po dokonaniu modyfikacji nagrania. Stopień poprawy jest różny w zależności od typu zakłócenia oraz stopnia podniesienia częstotliwości podstawowej.

Jak widać na wykresach przedstawionych na rysunku 7 dla szumu różowego znacząca poprawa współczynnika MOS nastąpiła dla małych lub ujemnych SNR. Jeśli SNR jest duży (większy od zera), zmiana jest znikoma. Dla sygnału *babble speech* sytuacja jest inna – poprawa następuje w zasadzie w każdym zakresie SNR.



Rys.7. Poprawa wskaźnika PESQ MOS dla różnego stopnia podniesienia F0 w zależności od szumu towarzyszącego nagraniom. Na osi poziomej SNR [dB] dla sygnału *babble speech* (pierwszy wykres) oraz dla szumu różowego (drugi wykres) miksowanego z nagraniami, na osi pionowej wyliczony wskaźnik PESQ

Wyniki liczbowe porównań przedstawiono w tab. 2 i 3.

Tabela 2. Porównanie współczynników PESQ MOS dla wykonanych nagrań i różnego poziomu zmiksowanego szumu różowego

Poziom SNR zmiksowanego szumu różowego [dB]	Nagranie w ciszy – MOS	F0 zwiększone o 40% - MOS	F0 zwiększone o 50% – MOS
-10	0,7428	1,0611	1,0920
-5	0,7403	1,1149	1,2333
0	0,9462	1,0982	1,0861
5	1,3131	1,4041	1,3384
10	1,6565	1,6741	1,6864

Tabela 3. Porównanie współczynników PESQ MOS dla wykonanych nagrań i różnego poziomu zmiksowanego *babble speech*

Poziom SNR zmiksowanego *babble speech* [dB]	Nagranie w ciszy – MOS	F0 zwiększone o 40% - MOS	F0 zwiększone o 50% - MOS	F0 zwiększone o 60% – MOS
-10	1,0234	1,3115	1,6047	1,6545
-5	1,2769	1,5105	1,9232	1,9158
0	1,6855	1,9500	2,2429	2,2513
5	1,9905	2,3161	2,6058	2,6043
10	2,3313	2,6335	2,9631	2,9606

7. WNIOSKI

Przeprowadzone eksperymenty wykazały, że istnieje możliwość poprawy jakości mowy, a co za tym idzie zwiększenia jej subiektywnej zrozumiałości poprzez wykonanie relatywnie prostych operacji na sygnale wejściowym – tj. poprzez zwiększenie częstotliwości podstawowej sygnału oraz zwiększenie natężenia dźwięku. Tego typu algorytm może znaleźć zastosowanie w tzw. alternatywnych urządzeniach poprawiających słuch lub w aparatach słuchowych [19], jak również w obszarze automatycznego rozpoznawania mowy (ASR – *Automatic Speech Recognition*) [20].

Wykazano możliwość poprawy estymacji współczynnika *Mean Opinion Score*, a nie jego realnej wartości, która może zostać potwierdzona jedynie w badaniach subiektywnych. Tym niemniej algorytm PESQ MOS stosowany jest w telekomunikacji równolegle z testami odsłuchowymi bądź niezależnie ze względu na fakt, iż wyniki estymacji w procesie PESQ są skorelowane z wynikami badań subiektywnych w ok. 93%. Można więc założyć, że wyniki eksperymentów potwierdzonych pomiarem PESQ mogą być w dużej mierze zbieżne z potencjalnym rezultatem badań subiektywnych.

Wykonane modyfikacje sygnału są łatwe do uzyskania i mogą być z powodzeniem stosowane w systemach czasu rzeczywistego, co umożliwia zastosowanie ich w systemach działających w hałasie, np. w protezach słuchu lub w systemach ostrzegania o zagrożeniach. Warto zauważyć, że dodatkowym aspektem, związanym z mową lombardzką, jest problem automatycznego rozpoznawania mowy w warunkach hałasu. Wartości parametrów dla sygnałów nie zawierających zakłócenia i nagranych/powstałych w warunkach hałasu (a więc z efektem Lombarda) będą uzyskiwać inne wartości, co przekłada się na efektywność rozpoznawania [21][22][23].

Mowa lombardzka zawiera znacznie więcej charakterystycznych cech, m.in. dłuższy czas trwania samogłosek, zwiększenie nachylenia widma, przesunięcie formantów w skali częstotliwości, itd. Dlatego, w kolejnych eksperymentach wprowadzone zostaną modyfikacje tych parametrów, co powinno w większym stopniu poprawić obiektywne wskaźniki jakości mowy.

Jak wspomniano wcześniej, w ramach wykonanych badań przygotowano zestaw zdań nagranych z różną prozodią dla głosów żeńskich i męskich. Dlatego aspekt ekspresji prozodii będzie tematem dalszych badań. Ponadto, ponieważ w literaturze można spotkać sprzeczne doniesienia dotyczące, w jakim stopniu efekt Lombarda zależy od płci [24][25][26], dlatego przewidywane jest również przeprowadzenie analiz porównawczych dla wypowiedzi kobiet i mężczyzn w warunkach hałasu.

Praca powstała częściowo w wyniku realizacji projektu badawczego o nr DEC-2015/17/B/ST6/01874
finansowanego ze środków Narodowego Centrum Nauki.

8. BIBLIOGRAFIA

1. Lombard E., Le signe de l'élévation de la voix (translated from French), Ann. des Mal. l'oreille du larynx, vol. 37, no. 2, pp. 101–119, 1911.
2. Lu Y., Cooke M., Speech production modifications produced by competing talkers, babble, and stationary noise, Journal of the Acoustical Society of America, 124, 2008, 3261–3275.

3. Kleczkowski P., Żak A., Król-Nowak A., Lombard Effect in Polish Speech and its Comparison in English Speech, Archives of Acoustics, vol. 42, no. 4, pp. 561–569, 2017, doi: 10.1515/aoa-2017-0060.
4. Boril H., Fousek P., Höge H., Two-Stage System for Robust Neutral/Lombard Speech Recognition, Interspeech, 2007.
5. Therrien A. S., Lyons J., Balasubramaniam R., Sensory Attenuation of Self-Produced Feedback: The Lombard Effect Revisited, PLoS One, vol. 7, no. 11, 2012.
6. Zollinger S.A., Brumm H., The evolution of the Lombard effect: 100 years of psychoacoustic research, Behaviour, 148, 2011, 1173–1198.
7. Bapineedu G., Analysis of Lombard effect speech and its application in speaker verification for imposter detection, Language Technologies Research Centre, International Institute of Information Technology.
8. Lau P., The Lombard Effect as a Communicative Phenomenon, UC Berkeley Phonology Lab Annual Report, 2008.
9. Junqua J.-C., Fincke S., Field K., The Lombard effect: a reflex to better communicate with others in noise, 1999 IEEE Int. Conf. Acoust. Speech, Signal Process. Proceedings. ICASSP99 (Cat. No.99CH36258), pp. 2083–2086 vol. 4, 1999.
10. Whitepaper PESQ: An Introduction, Psytechnics Limited, 2001.
11. Single-ended method for objective speech quality assessment in narrow-band telephony applications, ITU-T Recommendation P.563, 2004.
12. ITU-T. Methods for subjective determination of transmission quality. Recommendation P.800, Aug. 1996.
13. ITU-T. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs. Recommendation P.862, Feb. 2001.
14. Nishiura T., Detection for Lombard speech with second-order mel-frequency cepstral coefficient and spectral envelope in beginning of talking speech, The Journal of the Acoustical Society of America, 2013.
15. ITU-T. Mean opinion score (MOS) terminology. Recommendation P.800.1, July 2006.
16. ITU-R BS.1284: ogólne metody subiektywnej oceny jakości dźwięku.
17. ITU-R BS.1116: ocena małych zniekształceń dźwięku (test ABC).
18. ITU-T. Mapping function for transforming P.862 raw result scores to MOS-LQO. Recommendation P.862.1, Nov. 2003.
19. Poremski T., Szymański P., Kostek B., Aparat słuchowy a alternatywne urządzenia poprawiające słyszenie, Otorynolaryngologia 2018, 17(2): 49-56, www.mediton.pl/orl.
20. Marxer, R. Barker J. Alghamdi N., The impact of the Lombard effect on audio and visual speech recognition systems, Speech Communication, vol. 100, pp. 58-68, June 2018, https://doi.org/10.1016/j.specom.2018.04.006.
21. Boril H., Fousek P., Sündermann D., Cerva P., Zdansky J., Lombard Speech Recognition: A Comparative Study, InterSpeech 2007.
22. Boril H., Pollák P., Design and Collection of Czech Lombard Speech Database, http://www.isca-speech.org/archive/interspeech_2005/i05_1577.html.
23. Vlaj D., Kacic Z., The Influence of Lombard Effect on Speech Recognition in: Speech Technologies, Chapter 7, pp. 151-168.
24. Egan J. P., Psychoacoustics of the Lombard voice response, J. Aud. Res. 12, 1972, 318–324.
25. Zollinger S. A., Brumm H., The Lombard effect, Curr. Biol., vol. 21, no. 16, pp. R614–R615, 2011.
26. Stowe L. M., Golob E. J. Evidence that the Lombard effect is frequency-specific in humans. The Journal of the Acoustical Society of America, 134(1):640-647, 2013, doi:10.1121/1.4807645.

A STUDY ON IMPROVING OBJECTIVE QUALITY INDICATORS OF SPEECH UTTERANCES IN NOISE CONDITIONS

The aim of the work is to modify the speech signal in order to improve objective speech quality indicators after mixing the useful signal with noise or with an interfering signal. Modifications made to the signal are based on the characteristics of the Lombard speech, and in particular on the effect of raising the fundamental frequency F0. The recording session included sets of words and sentences in Polish, recorded in silence, as well as in the presence of interfering signals, i.e. pink noise and so-called babble speech, also referred to as the "cocktail-party" effect. As a part of the research, speech samples were processed - both sentences and words spoken by men. The study shows that raising the fundamental frequency results in increased values of the speech quality index, measured using the PESQ (Perceptual Evaluation of Speech Quality) standard.

Keywords: Lombard effect; PESQ (*Perceptual Evaluation of Speech Quality*); speech parameters.