Sposób i układ do poprawy jakości sygnału mowy w systemach rozpoznawania mowy i komunikacyjnych

Przedmiotem wynalazku jest sposób i układ do poprawy jakości sygnału mowy, zwłaszcza w systemach automatycznego rozpoznawania mowy, przeznaczony zwłaszcza do konferencji multimedialnych.

W znanych urządzeniach elektronicznych powszechnym sposobem komunikacji człowieka z urządzeniem jest komunikacja głosowa. Istniejące metody rozpoznawania mowy, mówcy oraz komend głosowych osiągają wysoką skuteczność działania w dobrych warunkach akustycznych. W przypadku obecności w otoczeniu dźwięków niepożądanych rozpoznawanie mowy staje się poważnym wyzwaniem. Przykładowe scenariusze użycia, w których występują silne zakłócenia to korzystanie z urządzenia na ulicy, lotnisku, w samolocie lub sterowanie komputerem pokładowym w samochodzie. Czynnikami istotnie pogarszającymi jakość sygnału akustycznego są zwłaszcza zakłócenia addytywne i obecność innych źródeł mowy. Poszukiwane są sposoby poprawy jakości sygnału mowy, mające na celu eliminację negatywnego wpływu zakłóceń, obniżających skuteczność rozpoznawania mowy, mówcy lub ogólnie zrozumiałość sygnału mowy. Jednym z istotnych trendów jest dołączenie do informacji pochodzącej z modalności akustycznej danych pozyskanych z modalności wizyjnej. Sygnał wizyjny nie ulega degradacji w wyniku wyżej wymienionych czynników pogarszających jakość sygnału fonicznego.

Z opisu patentowego US6594629 znane jest rozwiązanie dotyczące automatycznego rozpoznawania mowy z wykorzystaniem modalności wizyjnej oraz fonicznej. Kamera wizyjna wykorzystywana jest do detekcji aktywności głosowej, estymacji pozy twarzy oraz ekstrakcji dodatkowych informacji o artykułowanych słowach. Pojedynczy mikrofon wykorzystany jest do parametryzacji mowy oraz detekcji aktywności głosowej.

Z opisu patentowego US6314396 znane jest urządzenie do automatycznego kontrolowania wzmocnienia w torze fonicznym

wykorzystywanym w systemach automatycznego rozpoznawania mowy. Na podstawie parametrów sygnału fonicznego następuje detekcja sygnału tożsamego z mową, następnie zgodnie z wynikiem detekcji następuje normalizacja sygnału w celu wytłumienia fragmentów sygnału nie będącego mową oraz wzmocnienie fragmentów mowy. Tak przetworzony sygnał przekazywany jest do modułu rozpoznawania mowy.

Z opisu patentowego US6731334 znany jest układ poprawy jakości mowy składający się z macierzy mikrofonów, które nakierowywane są na aktywnego mówcę. Proponowane wykorzystanie systemu to wideokonferencje z wieloma użytkownikami. Dzięki wykorzystaniu informacji z macierzy mikrofonów możliwe jest ograniczenie ilości kamer poprzez nakierowywanie ich pozycji na aktywnego mówcę.

Z opisu patentowego US8761412 znane jest urządzenie do poprawy jakości sygnału fonicznego z wykorzystaniem kamery wizyjnej oraz macierzy mikrofonów. Działanie układu polega na lokalizacji mówcy na podstawie analizy obrazu z kamery, następnie ta informacja przekazywana jest w celu odpowiedniego nastawienia parametrów macierzy mikrofonów, tak aby główna wiązka charakterystyki kierunkowej tej macierzy mikrofonowej była skierowana w stronę mówcy.

Z opisu patentowego US7684982 znane jest urządzenie wykorzystujące kamerę oraz mikrofon w celu poprawy jakości rejestrowanego sygnału mowy. Za pomocą analizy obrazu z kamery lokalizowana jest twarz oraz aktywność ust, tożsama z aktywnością głosową mówcy. Fragmenty sygnału odpowiadające mowie znajdowane są z wykorzystaniem parametrów ekstrahowanych z sygnału fonicznego oraz wizyjnego. Fragmenty sygnału niezawierające mowy wykorzystywane są do pobierania próbek sygnału tła akustycznego oraz szumu panującego w środowisku. Pozyskane próbki z szumem wykorzystywane są do filtracji sygnału mowy poprzez odejmowanie widmowe sygnału zarejestrowanego i zbuforowanych próbek szumowych.

Sposób poprawy jakości sygnału mowy w systemach rozpoznawania mowy i komunikacyjnych, który odebrany jest z odbiornika dźwięku, polegający na rejestrowaniu twarzy użytkownika przez odbiornik wizyjny i generowaniu sygnałów lokalizujących jego twarz oraz aktywność ust, charakteryzuje się tym, że w układzie detekcji twarzy implementuje się pierwszy algorytm realizujący zadanie lokalizacji twarzy w sygnale wizyjnym. W układzie detekcji ust implementuje się drugi algorytm realizujący wykrywanie aktywności ust w sygnale wizyjnym obrazującym obszar twarzy oraz implementuje się trzeci algorytm realizujący pomiar aktywności ust w oparciu o estymatę konturu ust wokół wewnętrznych krawędzi górnej i dolnej wargi. Do odbiornika dźwięku, korzystnie macierzy mikrofonów lub akustycznej sondy natężeniowej, podłącza się układ filtracji kierunkowej, w którym implementuje się czwarty algorytm realizujący filtrację przestrzenną. Sygnał zawierający informację o lokalizacji twarzy przesyła się z układu detekcji twarzy do układu filtracji kierunkowej, a sygnał zawierający informację o aktywności ust przesyła się z układu detekcji ust do modulatora amplitudowego. Odebrany sygnał mowy z odbiornika dźwięku poddaje się filtracji przestrzennej w układzie filtracji kierunkowej. Przefiltrowany sygnał mowy kieruje się do modulatora amplitudowego, w którym ingeruje się w wartość amplitudy przesłanego przefiltrowanego sygnału mowy, w ten sposób, że fragmenty czasowe niezawierające sygnału mowy zeruje się, a jednocześnie wzmacnia się korzystnie przefiltrowany sygnał mowy rejestrowany w momentach czasu, w których przy pomocy trzeciego algorytmu wykrywa się aktywności ust, po czym uzyskany zmodyfikowany sygnał mowy, przekazuje się do modułu automatycznego rozpoznawania mowy.

Korzystnie z układu filtracji przestrzennej przefiltrowany sygnał mowy kieruje się do układu eliminacji zakłóceń występujących w środowisku akustycznym. W układzie eliminacji zakłóceń, z uzyskanego przefiltrowanego sygnału mowy, przy pomocy zaimplementowanego piątego algorytmu z wykorzystaniem sygnału zawierającego informację o aktywności ust, wybiera się przedziały czasu, w których nie wykrywa się aktywności ust i określa się widmo szumu występującego w tych przedziałach czasu. Profil szumu obliczany jest w pasmach częstotliwościowych, korzystnie z wykorzystaniem transformaty

Fouriera. Następnie usuwa się zaszumienie sygnału użytecznego z wykorzystywaniem widmowego odejmowania szumu od zmodyfikowanego sygnału mowy lub za pomocą filtracji adaptacyjnej. Przefiltrowany sygnał mowy kieruje się do modulatora amplitudowego, przy pomocy którego usuwa się artefakty dźwiękowe powstałe z fragmentów sygnału niezawierających mowy po filtracji przestrzennej oraz widmowej.

Układ do poprawy jakości sygnału mowy w systemach rozpoznawania mowy i komunikacyjnych, składający się odbiornika dźwięku połączonego z układem filtracji przestrzennej oraz odbiornika wizyjnego połączonego z układem detekcji twarzy, który połączony jest z układem filtracji przestrzennej charakteryzuje się według wynalazku tym, że układ detekcji twarzy połączony jest z układem detekcji ust, który połączony jest z modulatorem amplitudowym, który połączony jest z układem filtracji przestrzennej.

Korzystnie układ filtracji przestrzennej połączony jest z modulatorem amplitudowym poprzez układ regulacji zakłóceń, który połączony jest z układem detekcji ust.

W sposobie według wynalazku zastosowany został sposób oddziaływania na sygnał akustyczny oparty na próbkowaniu natężenia dźwięku w odróżnieniu od przestrzennego próbkowania ciśnienia akustycznego. W sposobie tym wprowadzona została filtracja przestrzenna sygnału w dziedzinie częstotliwości na bazie sygnałów z wektorowego czujnika akustycznego. Wykorzystanie sensora głębi i przetwarzanie obrazu trójwymiarowego, w odróżnieniu od znanych rozwiązań bazujących na sygnale dwuwymiarowym przyczyniło się do podwyższenia jakości filtracji sygnału mowy.

W wyniku wykorzystania wynalazku uzyskuje się podwyższoną jakość sygnału mowy w zakresie dynamiki, zwłaszcza odstępu sygnału użytecznego od zakłóceń.

Wynalazek opisany jest bliżej w przykładach wykonania i na rysunku, na którym fig. 1 przedstawia schemat blokowy sposobu, a fig. 2 schemat blokowy rozwinięcia sposobu.

Przykład 1

Jak pokazano na fig.1, układ do poprawy jakości sygnału mowy składa się z odbiornika dźwięku OD w postaci macierzy mikrofonów, połączonego z układem filtracji przestrzennej UF oraz odbiornika wizyjnego K połączonego z układem detekcji twarzy UT, który połączony jest z układem filtracji przestrzennej UF.

Układ detekcji twarzy UT połączony jest z układem detekcji ust US, który połączony jest z modulatorem amplitudowym MA, który połączony jest z układem filtracji przestrzennej UF.

W układzie detekcji twarzy UT implementuje się pierwszy algorytm AL1 realizujący zadanie lokalizacji twarzy w sygnale wizyjnym W1.

W układzie detekcji ust US implementuje się drugi algorytm AL2 realizujący wykrywanie aktywności ust w sygnale wizyjnym W1 obrazującym obszar twarzy oraz implementuje się trzeci algorytm AL3 realizujący pomiar aktywności ust w oparciu o estymatę konturu ust wokół wewnętrznych krawędzi górnej i dolnej wargi,

Do odbiornika dźwięku OD w postaci sondy natężeniowej, podłącza się układ filtracji kierunkowej UF, w którym implementuje się czwarty algorytm AL4 realizujący filtrację przestrzenną.

W trakcie rejestracji przez odbiornik wizyjny K obrazu zawierającego twarz, przy pomocy układu detekcji ust US wyznacza się położenie ust, a następnie wewnątrz tego samego układu detekcji ust US wyznacza się aktywność ust. Sygnał zawierający informację o lokalizacji twarzy Z1 przesyła się do układu filtracji kierunkowej UF, a sygnał zawierający informację o aktywności ust Z2 przesyła się do modulatora amplitudowego MA.

Odebrany sygnał mowy S1 z macierzy mikrofonów OD poddaje się filtracji przestrzennej w układzie filtracji kierunkowej UF. Filtracja przestrzenna dokonywana jest w dziedzinie częstotliwości poprzez obliczenie widmowej dystrybucji kątów natarcia dźwięku w płaszczyźnie azymutu i elewacji. Następnie dla danego pasma częstotliwościowego sygnału odebranego przez układ odbiornika dźwięku OD o częstotliwości środkowej f sygnał jest przetwarzany według wzoru:

$$S_2(f) = S_1(f) \cdot w(f)$$

gdzie $S_2(f)$ to sygnał w paśmie wokół częstotliwości f po filtracji przestrzennej, $S_2(f)$ to sygnał odebrany, a w(f) to funkcja ważąca przykładowo dana wzorem:

$$w(f) = 0.5 + 0.5\cos[\varphi(f) - \alpha]$$

gdzie $\varphi(f)$ to kąt padania fali dźwiękowej o częstotliwości w paśmie o częstotliwości środkowej f w płaszczyźnie azymutu, a kąt α wyznacza kierunek, na którym wykryto mówcę na podstawie analizy sygnału wizyjnego.

Uzyskany przefiltrowany sygnał mowy S2 kieruje się do modulatora amplitudowego MA, w którym ingeruje się w wartość amplitudy przesłanego przefiltrowanego sygnału mowy S2, w ten sposób że fragmenty czasowe niezawierające sygnału mowy zeruje się, a jednocześnie wzmacnia się przefiltrowany sygnał mowy S2 rejestrowany w momentach czasu, w których przy pomocy trzeciego algorytmu AL3 wykrywa się aktywność ust.

Uzyskany zmodyfikowany sygnał mowy SF, przekazuje się do układu rozpoznawania mowy UM.

W celu poprawy jakości sygnału mowy w systemie rozpoznawania mowy i komunikacyjnym z wykorzystaniem metody foniczno-wizyjnej, użytkownik lub użytkownicy muszą być widoczni w polu widzenia kamery w czasie, kiedy formułowana jest wypowiedź. Charakter pracy układu pozwala na wykorzystanie systemu w szumowych warunkach akustycznych, przykładowo gdy w środowisku występują zakłócenia pochodzące od rozmów innych osób lub hałas pochodzący od samochodu, szum przemysłowy lub inne, jak również w warunkach wolnych

od zakłóceń akustycznych. Układ będący przedmiotem wynalazku w sposób automatyczny zwiększa czułość układu mikrofonów na kierunku osoby aktualnie mówiącej, poprzez wykorzystanie sprzężenia z detektorem aktywności ust, który przekazuje informację o pozycji aktywnego mówcy. Sygnał mowy użytkownika doprowadzany może zostać do układu rozpoznawania mowy UM lub w formie ulepszonej zostać przesłany do sieci komputerowej SK i odsłuchany przez współuczestników zdalnej konferencji multimedialnej. W przypadku przesłania do układu rozpoznawania mowy UM rozpoznaje wypowiedzi w bliskim odstępie czasu od chwili, w której mówca skończy je wypowiadać lub w czasie rzeczywistym, natomiast wynik rozpoznawania może być przesłany jako tekst do sieci komputerowej SK lub wyświetlony na ekranie.

Przykład 2

Układ zbudowany jest jak opisano w przykładzie 1, z tym, że układ filtracji przestrzennej UF połączony jest z modulatorem amplitudowym MA poprzez układ regulacji zakłóceń UEZ, który połączony jest z układem detekcji ust US.

Sposób poprawy jakości sygnału mowy realizuje się jak opisano w przykładzie 1, z tym, że z układu filtracji przestrzennej UF przefiltrowany sygnał mowy S2 kieruje się do układu eliminacji zakłóceń UEZ występujących w środowisku akustycznym. Z uzyskanego przefiltrowanego sygnału mowy S2, przy pomocy zaimplementowanego piątego algorytmu AL5 z wykorzystaniem sygnału zawierającego informację o aktywności ust Z2, wybiera się momenty czasu, w których nie wykrywa się aktywności ust i określa się widmo szumu.

Następnie usuwa się zaszumienie sygnału użytecznego z wykorzystywaniem widmowego odejmowania szumu od zmodyfikowanego sygnału mowy S2. Polepszony sygnał mowy S3 kieruje się do modulatora amplitudowego MA, przy pomocy którego usuwa się artefakty powstałe po filtracji przestrzennej w momentach czasu, w których brak dźwięków wydawanych przez użytkownika.

Prowadzenie konferencji multimedialnej realizuje się jak w przykładzie 1.