# Sterowanie systemami inteligentnego budynku z wykorzystaniem komunikacji głosowej

Ryszard Tadeusiewicz

**Streszczenie:** W artykule przedstawiono argumenty przemawiające za tym, że dla sterowania systemami technicznymi (a zwłaszcza informatycznymi) wchodzącymi w skład inteligentnego budynku bardzo korzystne jest stosowanie komunikacji głosowej. Wskazano zalety wykorzystania sygnału mowy zarówno przy komunikacji od systemów automatyki sterujących budynkiem do ludzi użytkujących te systemy, jak i komunikacji w przeciwną stronę, to znaczy od ludzi do sterowanych maszyn. O ile jednak zbudowanie systemu automatycznego powiadamiania ludzi (personelu obsługi budynku oraz użytkowników budynku) za pomocą syntetycznej mowy może być zrealizowane raczej łatwo i wygodnie, o tyle komunikacja w drugą stronę nastręcza wielu trudności. Generatory mowy syntetycznej są obecnie powszechnie dostępne, tanie i łatwe w użyciu. Dlatego w artykule tylko krótko wzmiankowano o zasadach ich budowy, nie zatrzymując na tym elemencie uwagi. Natomiast urządzenia do automatycznego rozpoznawania mowy są systemami o dużym stopniu komplikacji, a ich budowa i użytkowanie wymaga rozwiązania wielu problemów i pokonania wielu trudności. W pracy scharakteryzowano te trudności, a także krótko omówiono sposoby ich przezwyciężania, prowadząc w końcowej części artykułu do prezentacji całościowej koncepcji systemu automatycznego rozpoznawania mowy, mogącego znaleźć zastosowanie w sterowaniu systemami inteligentnego budynku.

**INTELLIGENT BUILDING CONTROL SYSTEMS USING VOICE**

***Abstract:*** *Paper presents advantages of the use of speech signal for communication between people and intelligent building control systems. In fact the list of advantages is long and include many items which together are worth efforts which are necessary when going to the practical applications of voice communication between man and machines in intelligent building. The communication under consideration can be realized in two directions. Easier but less useful is speech communication from machines to the people. This model of „automatic voice announcement" is very useful and easy for realization, because methods of automatic speech synthesis are good developed and available. Voice communication in opposite direction, e.g. form man to automatic system is much more complicated. In this case the system designer must solve several problems, selected and discussed in the paper. Nevertheless this effort should be done, because voice control is the best solution in many situations related to the intelligent building systems. In the paper general schema of speech recognition system is presented and discussed as well as some selected details of its realization are discussed for further use.*

## 1. Wprowadzenie

Systemy związane z techniką inteligentnych budynków, chociaż ich celem jest automatyzacja różnych czynności związanych z funkcjonowaniem infrastruktury technicznej budynku oraz z obsługą potrzeb ludzi znajdujących się w budynku, nigdy nie funkcjonują jako systemy całkiem niezależne od człowieka. Mimo zaawansowanej (najczęściej rozproszonej) automatyzacji, jaka jest tu stosowana, ludzie chcą móc wydawać polecenia systemom inteligentnego budynku i mają prawo wymagać, by ta komunikacja z systemami technicznymi była dla nich (dla ludzi!) maksymalnie wygodna. Wprawdzie sama funkcja sterowania różnymi obiektami, wykrywania zagrożeń i eliminacji zakłóceń przejmowana jest przez określone systemy monitoringu, regulatory, sterowniki, mikrokontrolery lub specjalizowane procesory, ale stawianie zadań tym układom automatyki, ustawianie ich parametrów albo przejmowanie kontroli w warunkach krytycznych – jest ciągle domeną człowieka. W związku z tym systemom automatyki stanowiącym wyposażenie współczesnych inteligentnych budynków stawia się określone wymagania także w zakresie ergonomii i wygody kontaktów z ludźmi – zarówno z użytkownikami inteligentnego budynku, jak i wchodzącymi w skład personelu obsługi budynku (automatyzacja nigdy nie eliminuje ludzkiego nadzoru w 100%).

W dawniejszych rozwiązaniach parametry i wartości zadane dla układów automatyki ustawiało się pracowicie na panelach sterujących z użyciem różnych kalibrowanych pokręteł, suwaków i innych ręcznie obsługiwanych nastawników. Potem przyszła epoka paneli wirtualnych, symulowanych za pomocą grafiki komputerowej, a także różnych ekranów dotykowych oraz urządzeń przenośnych wykorzystujących łączność bezprzewodową. Tak wygląda sfera komunikacji systemów automatyki z obsługującymi je ludźmi dziś i zapewne jeszcze przez kilka nadchodzących lat.

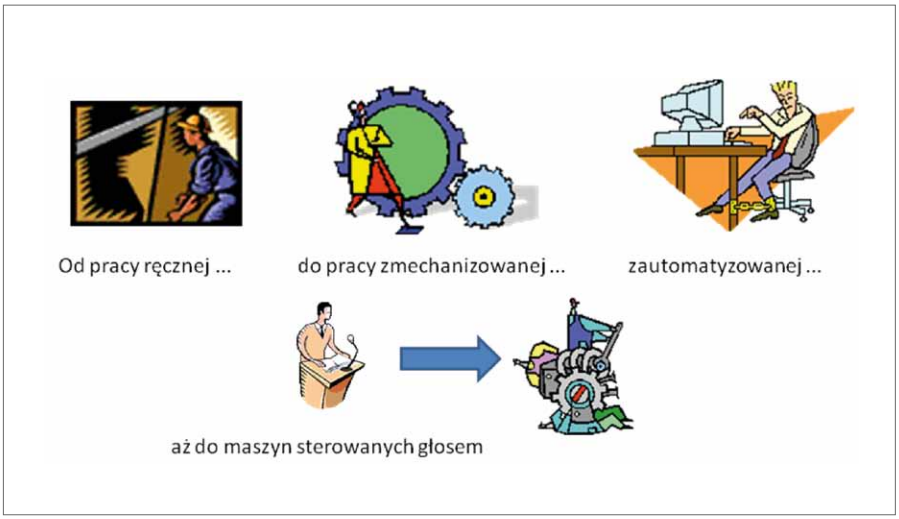
Warto jednak pomyśleć już dziś o rozwiązaniach technicznych, które pojawią się niebawem i mają szanse zrewolucjonizować technikę komunikacji między ludźmi a systemami automatyki. Takimi urządzeniami przyszłościowymi będą niewątpliwie systemy obsługiwane za pomocą sygnału mowy (rys. 1).

## 2. Dlaczego wybieramy do sterowania sygnał mowy?

Głosowe sterowanie maszyn i urządzeń ma wiele zalet w stosunku do rozwiązań stosowanych aktualnie, przy czym warto może wskazać na kilka z tych zalet, żeby uświadomić sobie, o jak dużą stawkę idzie. Wskażmy więc, że głosowe wprowadzanie poleceń do układów sterowania cechuje się między innymi następującymi walorami:

- stanowi formę komunikacji niewymagającą oddzielnego szkolenia, bo związaną z rodzajem ludzkim od najdawniejszych czasów (rys. 2);
- jest najwygodniejsze i najbardziej naturalne dla człowieka (rys. 3);
- nie wymaga fizycznego kontaktu z żadnym urządzeniem, może więc być realizowane na odległość, a także wtedy, gdy człowiek ma zajęte ręce (rys. 4);
- może być z powodzeniem realizowane przy braku widoczności (ciemność, zadymienie – rys. 5), a także w warunkach fizycznego obciążenia osoby wydającej polecenia;
- pozwala na wykorzystanie do celów sterowania powszechnie dostępnych systemów telefonii stacjonarnej i komórkowej (rys. 6);
- jest dostępne dla osób niepełnosprawnych (rys. 7).

Wymienione okoliczności sprawiają, że sterowanie za pomocą sygnału mowy jest bardzo dobrym rozwiązaniem we wszelkich systemach komunikacji człowieka z systemami technicznymi, a szczególnie przydatne może się okazać właśnie w inteligentnych budynkach, gdzie dzięki stałej rozbudowie nowych funkcji i nowych elementów infrastruktury rośnie także zakres możliwych (a czasem wręcz koniecznych) interakcji pomiędzy ludźmi i systemami automatyki, w jakie wyposażony jest budynek. Warto także zauważyć, że w odróżnieniu od sytuacji, w jakiej działają na przykład systemy automatyki przemysłowej, większość użytkowników systemów automatyki w inteligentnych budynkach stanowią ludzie, którzy nie są w żaden sposób specjalnie szkoleni w zakresie obsługi tych wszystkich urządzeń, do których mają dostęp. Komunikacja między tymi ludźmi a systemami technicznymi musi więc być szczególnie prosta i szczególnie intuicyjna (rys. 8) – a taka jest właśnie




**Rys. 1.** Sterowanie głosem jako logiczne następstwo rozwoju techniki sterowania

**Rys. 2.** Funkcję mowy posiadały prawdopodobnie nawet najdawniejsze humanoidy około 3,5 miliona lat temu

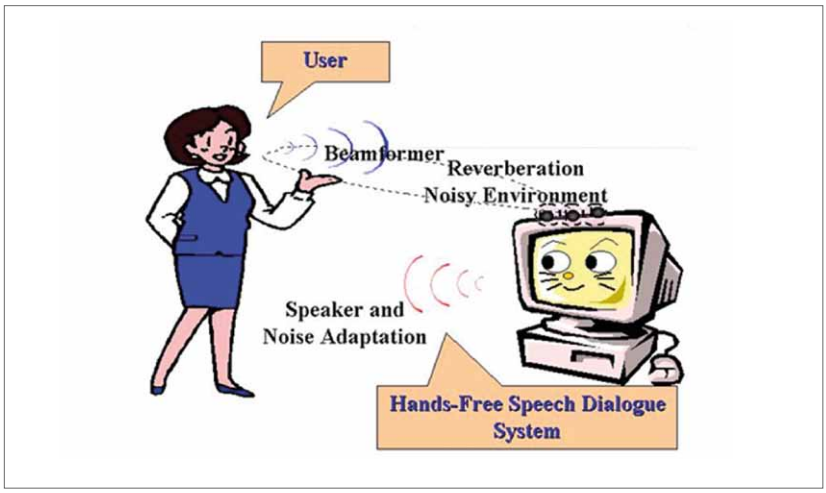
**Rys. 3.** Komunikacja głosowa jest dla ludzi najbardziej naturalna

**Rys. 4.** Sterowanie za pomocą mowy jest skuteczne, także gdy człowiek ma zajęte ręce

**Rys. 5.** Sterowanie głosowe nie wymaga oświetlenia

**Rys. 6.** Można przy sterowaniu wykorzystywać wszystkie elementy infrastruktury telefonicznej

**Rys. 7.** Sterowaniem głosowym mogą się posługiwać także osoby niepełnosprawne




**Rys. 8.** Uproszczone wyobrażenie systemu sterowania głosem

**Rys. 9.** Głosowe sterowanie różnymi urządzeniami ujawnia swoje zalety szczególnie w sytuacjach zagrożenia, gdzie ludzie wydający polecenia systemom muszą działać w stresie

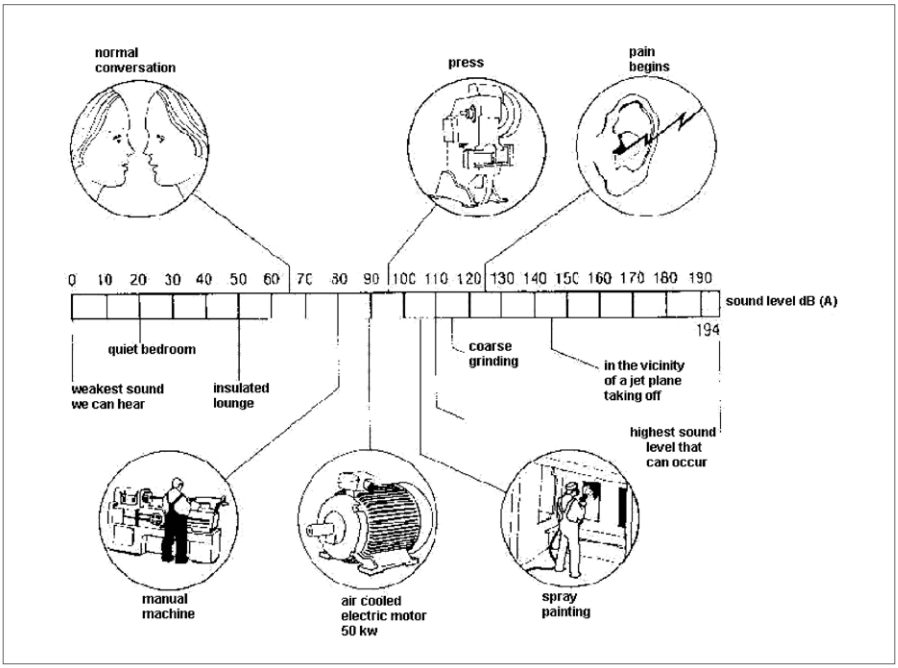
komunikacja z wykorzystaniem systemów sterowania głosowego.

Dodatkowo można wskazać, że w warunkach zagrożenia i stresu reakcja głosowa może być szybsza niż jakakolwiek manipulacja wymagająca użycia rąk i wiążąca się z odnalezieniem (wśród wielu innych) właściwego przycisku czy manipulatora, a także – jak wynika z badań psychofizycznych zdolności człowieka – jest w tych warunkach obarczona znacznie mniejszym prawdopodobieństwem popełnienia błędu (rys. 9).

Z tego powodu należy przewidywać i oczekiwać, że nowe rozwiązania automatyki w inteligentnych budynkach będą w coraz większym stopniu nastawione na wykorzystanie sygnału mowy jako nośnika informacji przy przekazywaniu poleceń człowieka kierowanych do systemu. Jest to tym bardziej naturalne, że sprawne systemy głosowej komunikacji w drugą stronę (to znaczy od systemu technicznego do człowieka z wykorzystaniem elektronicznych syntezatorów mowy) są już bardzo dobrze rozwinięte i szeroko spotykane – by wspomnieć tylko o wspomagających kierowców systemach nawigacji opartych na GPS. O ile jednak użycie sygnału mowy do przekazywania komunikatów od dowolnego systemu technicznego do człowieka jest zadaniem prostym i stosunkowo łatwym do realizacji, o tyle komunikacja w drugą stronę (głównie tu nas interesująca) – jest trudna i skomplikowana (rys. 10).

Spróbujemy teraz pokrótce wskazać, dlaczego tak trudna do technicznej realizacji jest głosowa komunikacja od człowieka do maszyny.


**Rys. 10.** Asymetria w głosowej komunikacji między człowiekiem i maszynami

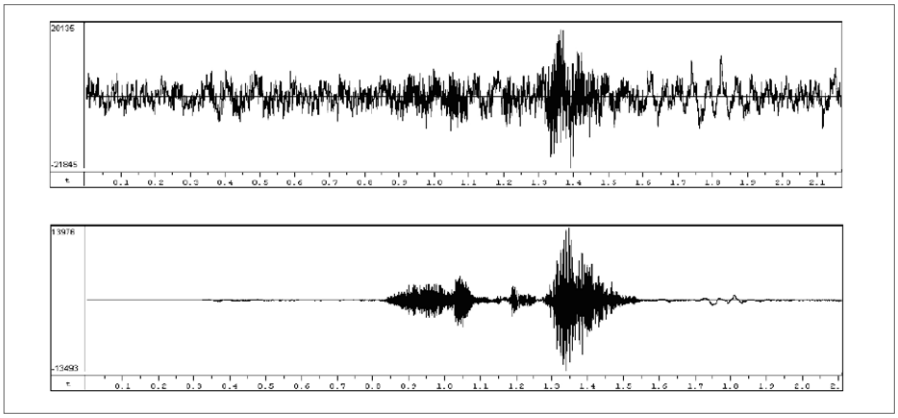



**Rys. 11.** Podlegający rozpoznawaniu sygnał mowy jest zwykle związany z różnymi sygnałami zakłócającymi

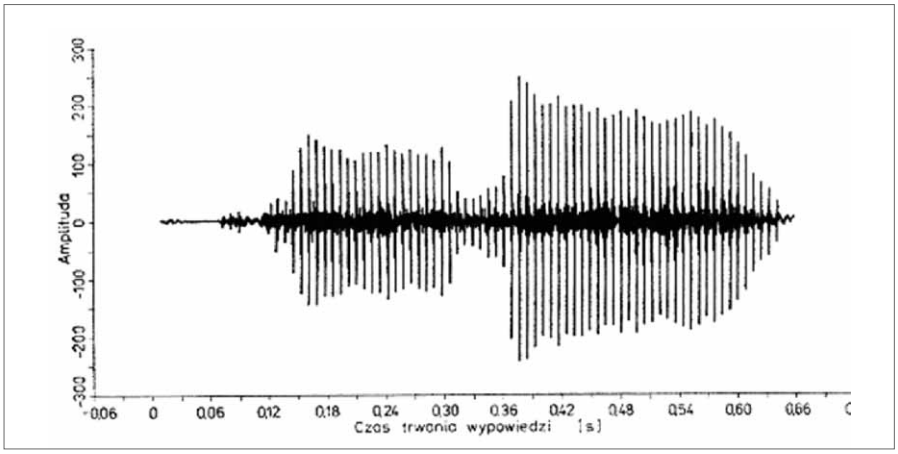
### 3. Sprawa najbardziej oczywista: zakłócenia

W systemach rozpoznawania mowy źródłem trudności są też sygnały akustyczne inne niż rozpoznawana mowa, dostające się na wejście systemu rozpoznającego (rys. 11). W najprostszym przypadku są to zakłócenia i szumy, które trzeba odfiltrować. Właściwie zadanie to wydaje się rutynowe i łatwe (w końcu filtrowane są przeróżne sygnały i każdy elektronik wie, jak to się robi), ale trzeba to zrobić naprawdę dokładnie i starannie (rys. 12), jeśli wynik rozpoznawania mowy ma być zgodny z naszymi oczekiwaniami.

Przy filtracji mającej na celu usunięcie zakłóceń domieszanych przez środowisko do analizowanego sygnału mowy można też usunąć pewne niekorzystne właściwości sygnału mowy leżące w samej jego naturze, ale także niekorzystnie wpływające na proces automatycznego rozpoznawania tego sygnału. Chodzi o niekorzystny skład spektralny sygnału mowy, w którym jest silna komponenta niskoczęstotliwościowa o stosunkowo małej wartości informacyjnej (składowe sygnału mowy o niskiej częstotliwości to głównie samogłoski, mające przeciętnie o blisko 20 dB większą energię, niż składniki spółgłoskowe), oraz niosąca niewielką energię składowa wysokoczęstotliwościowa, która jednak okazuje się krytyczna (wybitnie użyteczna) przy rozpoznawaniu znaczenia wypowiadanych słów. Zobrazowano to na rysunku 13, na którym widoczne są wysokie amplitudy sygnału w interwałach czasu odpowiadających samogłoskom – i prawie niewidoczne fragmenty przebiegu odpowiadające spółgłoskom.



**Rys. 12.** Sygnał mowy w formie, w jakiej jest zwykle rejestrowany w zastosowaniach praktycznych (silnie zakłócony – u góry), oraz ten sam sygnał zarejestrowany w warunkach laboratoryjnych (u dołu). Algorytmy pozwalające automatycznie rozpoznawać czyste sygnały (u dołu) zawodzą często przy zastosowaniu do sygnałów zakłóconych (u góry)




**Rys. 13.** Nierównomierne rozłożenie energii w sygnale mowy. Opis w tekście

Wiedząc o tym, że spółgłoski niosą z reguły więcej informacji o znaczeniu wypowiadanych słów niż samogłoski – zmierzamy przy automatycznej analizie sygnału mowy do tego, żeby ten mankament wyeliminować. Służy do tego specyficzny rodzaj filtracji sygnału mowy, zwany preemfazą. Charakterystyka częstotliwościowa filtra preemfazy podana jest na rysunku 14.

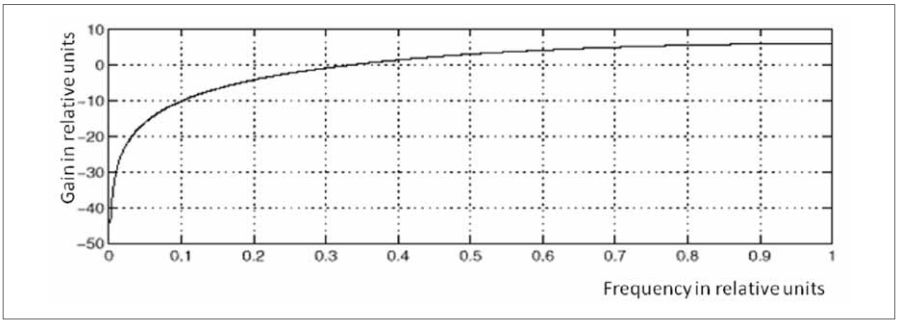
Dzięki zastosowaniu preeemfazy rozkład energii sygnału w paśmie wysokich i w paśmie niskich częstotliwości jest bardziej równomierny. Widać to na rysunku 15, przedstawiającym tę samą wypowiedź co na rysunku 13, ale po

procesie preemfazy. Dzięki preemfazie te części sygnału mowy, które odpowiadają spółgłoskom, zostają wzmocnione i uwypuklone. Ma to istotne znaczenie dla rozumienia mowy, bo spółgłoski odgrywają przy tym ważniejszą rolę niż samogłoski. Można tu przypomnieć fakt, że w niektórych językach bliskowschodnich przy zapisie słów rejestruje się tylko spółgłoski, co wystarcza do sprawnego czytania i rozumienia tekstu, natomiast samogłoski dodaje się podczas wypowiadania słów dla ułatwienia artykulacji oraz dla polepszenia percepcji sygnału.

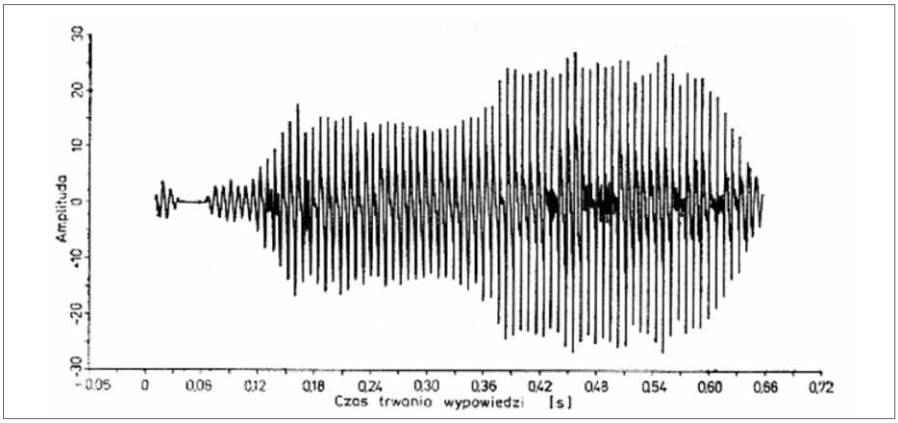
Opisane wyżej filtracje pozwalają usunąć zewnętrzne szumy pochwycone przez mikrofon wraz z sygnałem mowy oraz wewnętrzne właściwości samego sygnału mowy, utrudniające jego rozpoznanie. Znacznie trudniejszy problem techniczny pojawia się w sytuacji, gdy zakłóceniem dla analizowanego właśnie sygnału mowy jest… inny sygnał mowy (rys. 16).

Zagadnienie takie znane jest w literaturze jako tzw. *cocktail party problem*. Wymaga ono stosowania skomplikowanych metod tzw. dekonwolucji sygnału, które w tym miejscu tylko sygnalizujemy, nie rozwijając tego wątku.

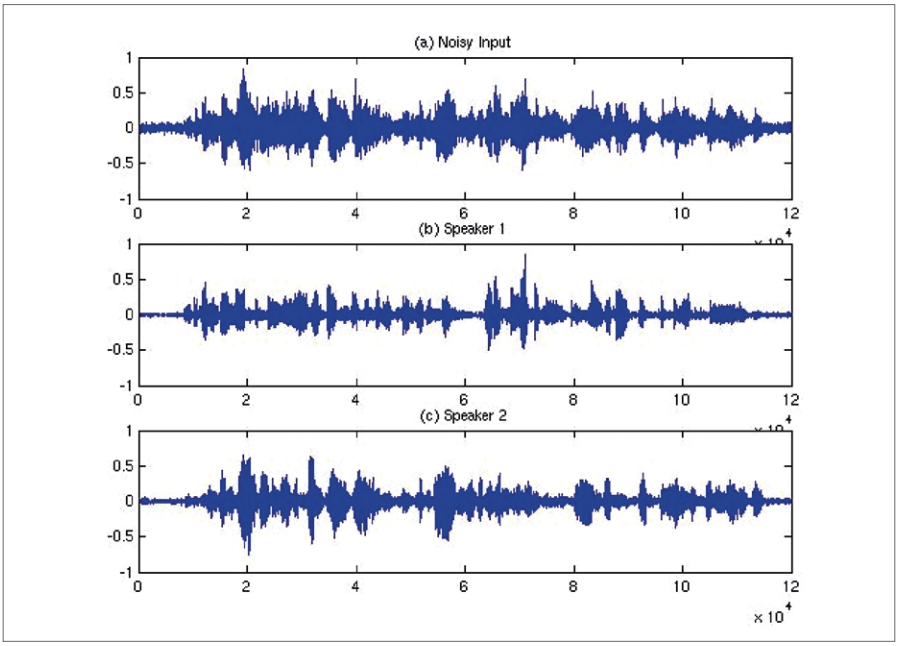
Na filtracji sygnału kłopoty z rozpoznawaniem sygnału mowy bynajmniej się nie kończą. Omówimy teraz kilka zagadnień związanych z bogatą zawartością sygnału mowy.




**Rys. 14.** Charakterystyka częstotliwościowa filtra preemfazy




**Rys. 15.** Sygnał mowy po procesie preemfazy



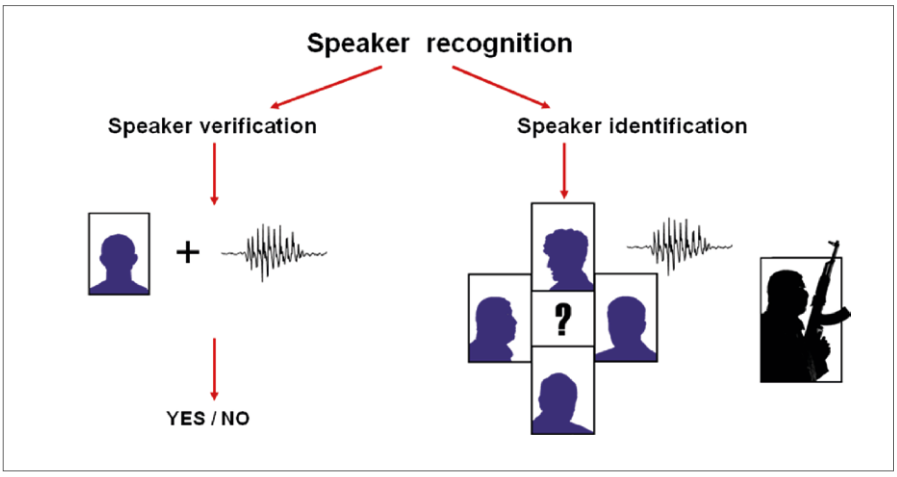

**Rys. 16.** Bardzo duże trudności przy automatycznym rozpoznawaniu mowy wiążą się z sytuacją równoczesnego mówienia przez kilku ludzi

## 4. Różne rodzaje informacji zawarte w sygnale mowy

Warto sobie uświadomić, że sygnał mowy niesie bardzo wiele różnego rodzaju informacji. Pierwsza i najbardziej oczywista z nich to informacja semantyczna. Analizując sygnał mowy za pomocą komputera, można więc próbować ustalić, jakie treści przekazała osoba mówiąca. Taka wiedza jest najbardziej użyteczna z punktu widzenia automatyki i na tym się dalej skoncentrujemy. Jednak sygnał mowy ma także inne komponenty, które w naszym zastosowaniu traktować będziemy jako zakłócenia, ale które w niektórych innych zastosowaniach mogą być głównym przedmiotem zainteresowania. Taką dodatkową informacją zawartą w sygnale mowy jest informacja osobnicza. Słysząc głos osoby mówiącej, możemy w wielu przypadkach łatwo ustalić, kto mówi. W głosie osoby mówiącej zawarta jest bowiem informacja o płci, wieku, a także innych cechach osobniczych mówcy. W automatyce to nam przeszkadza, bo chcemy (na ogół), żeby budowany system w taki sam sposób odbierał i interpretował docierające do niego polecenia i inne wypowiedzi – niezależnie od tego, kto wydał polecenie i jakie są indywidualne cechy głosu osoby mówiącej (rys. 17).

Są jednak zastosowania (wcale nie odległe od automatyki), w których głów-


**Rys. 17.** Sygnał mowy niesie informacje o płci, wieku i innych cechach indywidualnych mówcy



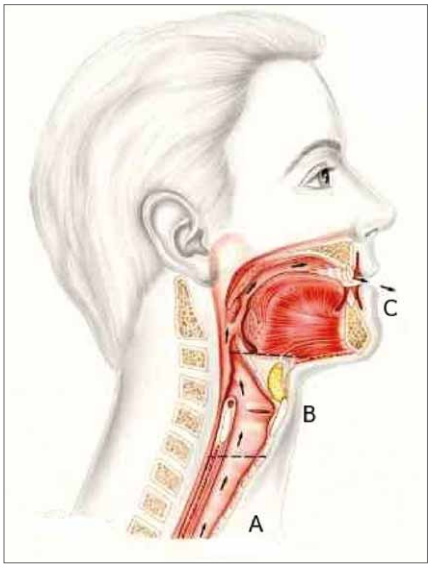

**Rys. 18.** Rozpoznawanie mowy może pozwalać na weryfikację lub identyfikację mówcy

nym celem analizy akustycznej sygnału mowy jest identyfikacja mówcy albo weryfikacja, czy jest on tym, za kogo się podaje (głos zamiast klucza, hasła albo PIK-kodu), co także może mieć zastosowanie w technice inteligentnych budynków (rys. 18), zwłaszcza że ma sporo zalet (rys. 19).

Wymienione dwa główne zastosowania automatycznego rozpoznawania mowy są wzajemnie przeciwstawne – w tym sensie, że wszystkie te cechy sygnału, które związane są z unikatowymi właściwościami głosów poszczególnych osób i pozwalają te osoby identyfikować, są jednocześnie źródłem problemów i kłopotów w momencie, kiedy chcemy zbudować algorytm rozpoznający treść wypowiedzi niezależnie od tego, kim jest osoba wypowiadająca określone słowa. I *vice versa*: Gdy chcemy zidentyfikować mówcę, to powinniśmy móc to uczynić niezależnie od tego, co on powiedział, bo nie zawsze mamy ten komfort, że możemy mieć wpływ na treść wypowiadanych słów, jak to było w znanej bajce o Sezamie. Jednak wtedy fakt, że dźwięki różnych głosek, sylab, słów i zdań są różne, co pozwala rozróżniać i rozpoznawać te wypowiedzi, stanowi przyczynę sporych trudności.

Warto dodać, że na rozpoznawaniu treści wypowiedzi oraz osoby mówiącej sprawa bynajmniej się nie kończy. Sygnał mowy niesie dodatkowo informację o stanie psychicznym. Wprawne ucho (oraz odpowiednio zaprogramowana aparatura) mogą dostarczyć informacji


**Rys. 19.** Zalety identyfikacji osób na podstawie brzmienia ich głosu są oczywiste




**Rys. 21.** W sygnale mowy odzwierciedlony jest stan zdrowia wszystkich narządów wchodzących w skład tzw. traktu głosowego

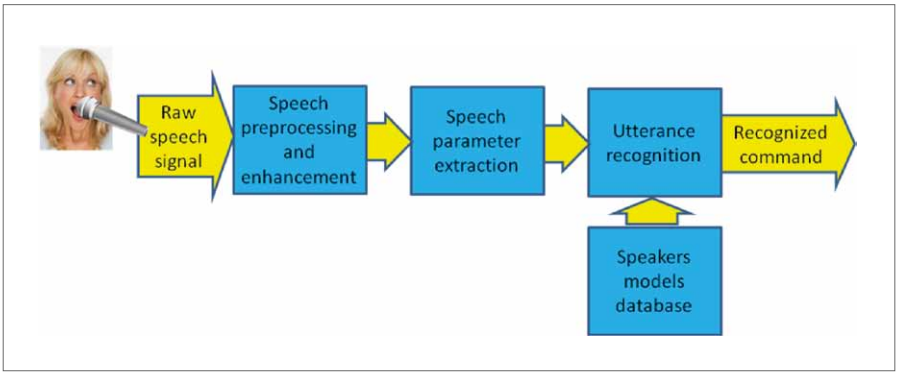

**Rys. 20.** Mowa zależy też od emocji osoby mówiącej

tym, w jakim nastroju jest osoba mówiąca. Można wykryć, że mówca jest smutny albo wesoły, że jest przestraszony albo podniecony, czasem także to, że kłamie albo nie mówi całej prawdy (rys. 20). W każdej ze wskazanych sytuacji sygnał mowy jest inny – nawet wtedy, gdy ta sama osoba wypowiada te same polecenia. Czasem to może być użyteczne w systemach automatyki (na przykład pozwala wykryć w głosie operatora to, że jest on zmęczony albo jest pijany), ale na ogół ta różnorodność przysparza kłopotów twórcom systemów automatycznego rozpoznawania mowy.

W sygnale mowy zawarte są też wiadomości o stanie zdrowia osoby mówiącej. Istnieje dziś obszerny dział inżynierii biomedycznej, który zajmuje się diagnostyką różnych chorób na podstawie charakterystycznych zjawisk akustycznych wykrywanych w sygnale mowy. Nie ułatwia to jednak zadania twórcom systemów sterowania kontrolowanych z pomocą komend wydawanych głosem – bo operator powinien mieć możliwość skutecznego wydania polecenia układowi automatyki także wtedy, gdy na przykład ma chrypkę po imprezie integracyjnej.

Nie chcąc całkowicie zniechęcić entuzjastów systemów sterowania kontrolowanych za pomocą mowy, nie wspomnimy o tym, że sygnał ten może zawierać informacje o pochodzeniu społecznym osoby mówiącej, o regionie kraju, w którym się wychowała (lub o tym, że mówca jest cudzoziemcem), a pośrednio także o wychowaniu i wykształceniu spikera.

Przedstawiając poniżej wybrane uwagi na temat techniki automatycznego rozpoznawania treści wypowiedzi, chcemy podkreślić, że ze względu na przeglądowy charakter tego opracowania nie będziemy w nim podawali żadnych szczegółów na temat tego, w jaki sposób dokonuje się aktualnie rozpoznawania mowy, jakim przekształceniom poddawany jest rejestrowany przez komputer sygnał akustyczny, jakie cechy fonetyczne są wydobywane w celu jego identyfikacji i klasyfikacji, a także jakie techniki automatycznego rozpoznawania są stosowane i z jakimi skutkami. Opracowanie, które by miało chociażby powierzchownie opisać wszystkie te zagadnienia, musiałoby mieć znacznie większą objętość i byłoby (ze względu na wysoce specjalistyczny charakter) słabo czytelne dla nieprzygotowanego czytelnika oraz źle osadzone w kontekście innych opracowań, składających się na te materiały naszej tradycyjnej konferencji.




**Rys. 22.** Podstawowe moduły wchodzące w skład systemu automatycznego rozpoznawania mowy

**Rys. 23.** Mikrofon zakładany na głowę – najlepsze źródło sygnału dla systemu automatycznego rozpoznawania mowy

## 5. Ogólne zasady budowy systemów sterowania głosowego

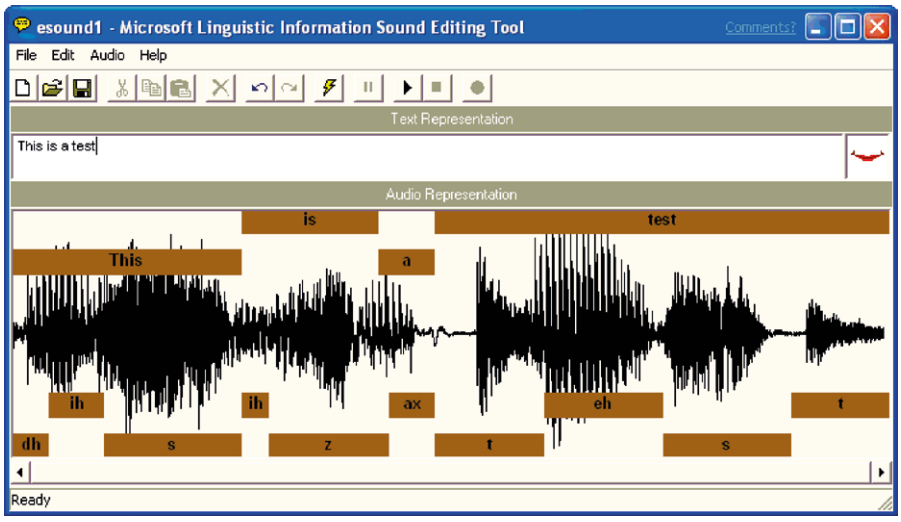
Ogólna struktura systemu przeznaczonego do sterowania za pomocą głosu przedstawiona jest na rys. 22.

Pierwszym elementem tego systemu jest mikrofon. Pozornie jest to urządzenie znane i łatwo dostępne. Jednak nie każdy mikrofon jest tak samo przydatny do celów budowy systemu automatycznego rozpoznawania mowy. Problemem, z którym się tu spotykamy, jest bowiem ogromna zmienność rejestrowanego sygnału mowy w zależności od odległości między ustami mówcy i mikrofonem. Dlatego tam, gdzie to jest możliwe, należy wykorzystywać mikrofony zakładane na głowę (rys. 23).

Sygnał mowy w takiej postaci, w jakiej rejestruje go mikrofon (rys. 24), jest jednak generalnie mało przydatny z punktu widzenia rozważanych tu systemów. Badania wykazują, że kształt i przebieg fali dźwiękowej rozpatrywanej w dziedzinie czasu bardzo silnie zmienia się w zależności od osoby mówcy, w zależności od tempa mowy, częściowo także w zależności od nastroju osoby mówiącej – natomiast w niewielkim stopniu odwzorowuje treść wypowiadanych poleceń.

Sygnał ten trzeba więc odpowiednio przetworzyć. Współczesna akustyka,

fonetyka i informatyka stwarzają łącznie bogaty zestaw narzędzi, które mogą być zastosowane do przetwarzania, analizy, rozpoznawania i rozumienia sygnału mowy [1]. Fakt ten ma doniosłe znaczenie w kontekście wielu zastosowań, bo ludzie niezwykle chętnie posługują się mową podczas komunikacji z innymi ludźmi (a czasem także w trakcie werbalizowania i porządkowania myśli na własny użytek), przeto systemy techniczne zdolne do odbierania i wykorzystywania naturalnego sygnału mowy człowieka mogą znaleźć wyjątkowo liczne i wyjątkowo użyteczne zastosowania praktyczne. Artykuł ten poświęcony jest dyskusji tych zastosowań technologii komputerowego przetwarzania mowy, które mogą być wykorzystane w automatyce, a zwłaszcza w technice inteligentnych budynków. Jest to oczywiście tylko pewien podzbiór zbiorowości wszystkich możliwych obecnie zastosowań technologii mowy, dlatego bardziej wymagających Czytelników odsyłamy do pozycji [2] bibliografii, w której zagadnienie możliwych zastosowań rozważanych tu metod technicznych przedstawione jest obszerniej i bardziej wyczerpująco. Dla głębszego wejścia w temat można także wykorzystać książkę [3], której pełny tekst jest dostępny w internecie, zaś osoby zainteresowane pierwszymi (najdawniejszymi, ale w znacznej części aktualnymi do dnia dzisiejszego) koncepcjami zastosowania sygnału mowy jako nośnika informacji w komunikacji między człowiekiem a systemami technicznymi mogą skorzystać z pozycji bibliograficznej [4].

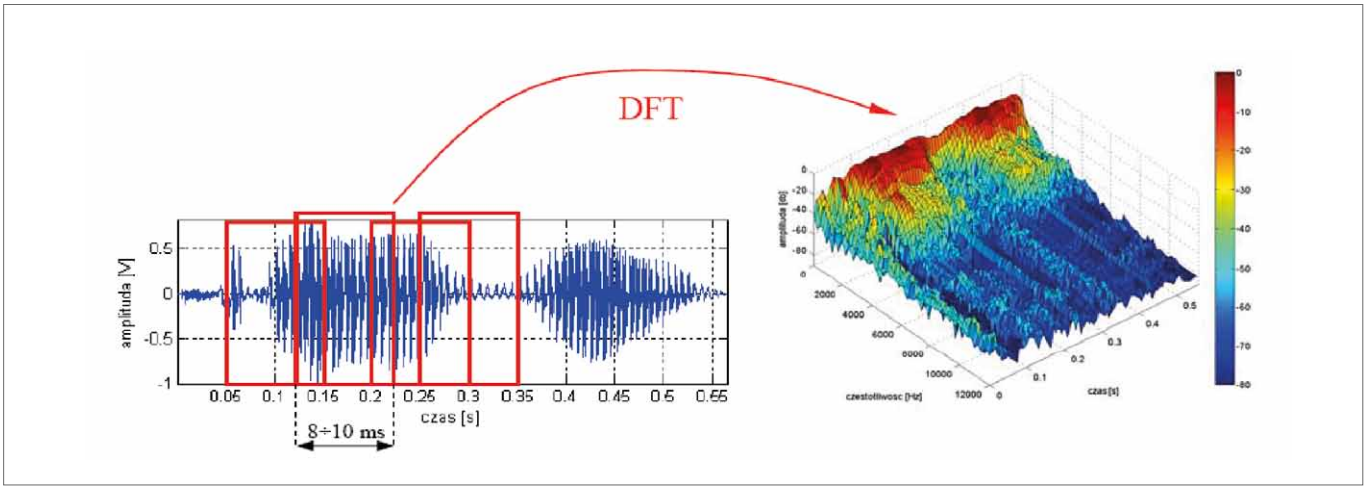



**Rys. 24.** Sygnał mowy w postaci przebiegu czasowego jest dla celów rozpoznawania treści wypowiedzi niemal całkowicie nieprzydatny

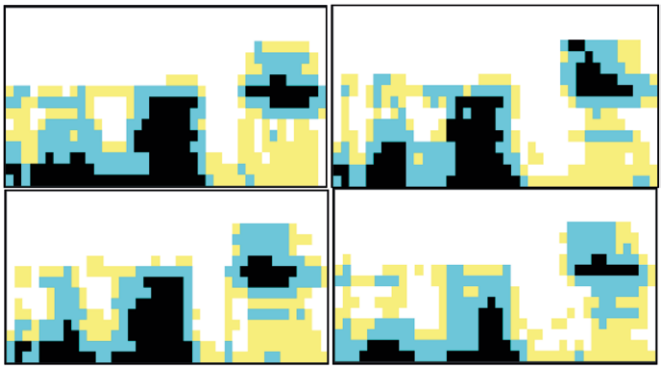
Wygodną podstawą do automatycznego rozpoznawania mowy są widma dynamiczne sygnału, tak zwane spektrogramy dynamiczne (rys. 25). Ich miłą cechą jest to, że można je stosunkowo łatwo uzyskać (dawniej z wykorzystaniem zestawu filtrów, dziś najczęściej przy użyciu takich transformat, jak DFT i FFT, a także przy pomocy przekształceń falkowych), a ponadto mowa reprezentowana w postaci spektrogramu dynamicznego może być stosunkowo łatwo rozpoznawana przy użyciu technik podobnych do tych, jakie są stosowane przy rozpoznawaniu obrazów (rys. 26).

Opisane do tej pory elementy pozwalają na przedstawienie najprostszej struktury systemu rozpoznawania mowy, który może znaleźć zastosowanie w automatyce inteligentnych budynków. System ten przedstawiony jest na rysunku 27, na którym rolę urządzenia wprowadzającego sygnał mowy do analizy pełni telefon komórkowy. Jest to bez wątpienia jedna z możliwości, ale w taki sam sposób będą działały pozostałe składniki systemu, jeśli w tym miejscu pojawi się dowolny inny mikrofon – na przykład rekomendowany mikrofon nagłowny (rys. 23).

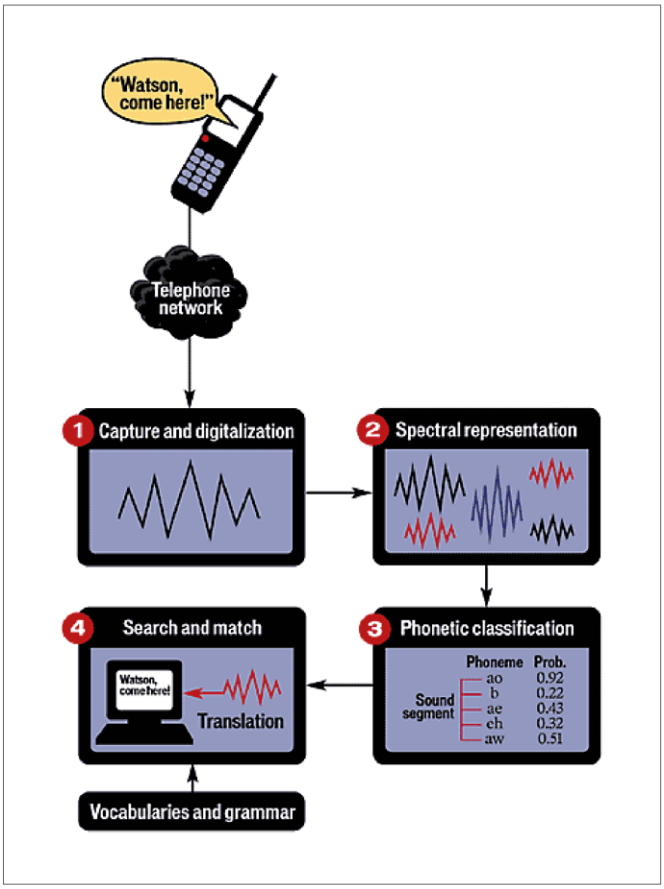
System z rysunku 27 oparty jest na zasadzie prostego porównywania aktualnie odebranego sygnału głosowego z zapamiętanymi wzorcami, w następstwie czego możliwe jest rozpoznanie pojedynczych słów czy nieskomplikowanych komend. Czynności przewidziane w tym systemie to (odwołując się do numerów bloków na rysunku):




**Rys. 25.** Sposób przetwarzania mowy poprzedzający zazwyczaj jej rozpoznawanie



**Rys. 26.** Proste elementy mowy (na przykład komendy) mogą być rozpoznawane na podstawie prostego podobieństwa całych spektrogramów. Cztery różne wypowiedzi słowa („koniec"). Widać wyraźne podobieństwa




**Rys. 27.** Prosty system rozpoznawania mowy, który bywa wykorzystywany w telefonii oraz w innych urządzeniach, którym dzięki temu można wydawać komendy głosowe

1. Pozyskanie sygnału mowy w postaci cyfrowej.
2. Transformacja sygnału do postaci widma dynamicznego (por. rys. 25).
3. Ocena podobieństwa wybranych segmentów rozpoznawanego sygnału do zapamiętanych wzorców. Rozpoznawane segmenty mogą mieć długość jednego, dwóch lub trzech fonemów (konkretnych realizacji głosek). Oceny podobieństwa wyraża się ilościowo.
4. Na podstawie sekwencji rozpoznanych segmentów wraz z przypisanymi im miarami pewności rozpoznania próbuje się rozpoznać słowa, frazy i całe zdania, posługując się specjalnie skonstruowanymi słownikami oraz bardzo uproszczonym modelem gramatyki.

Systemów o budowie omówionej wyżej powstało sporo, do różnych zastosowań, i w przypadku spełnienia kilku warunków dobrze sprawdzają się one w praktyce, umożliwiając głosowe sterowanie różnymi urządzeniami i różnymi funkcjami. Warunki, o których mowa, są jednak czasem trudne do spełnienia, gdyż dosyć istotnie ograniczają one działanie urządzenia. Wymieńmy i wskażmy, dlaczego są kłopotliwe.

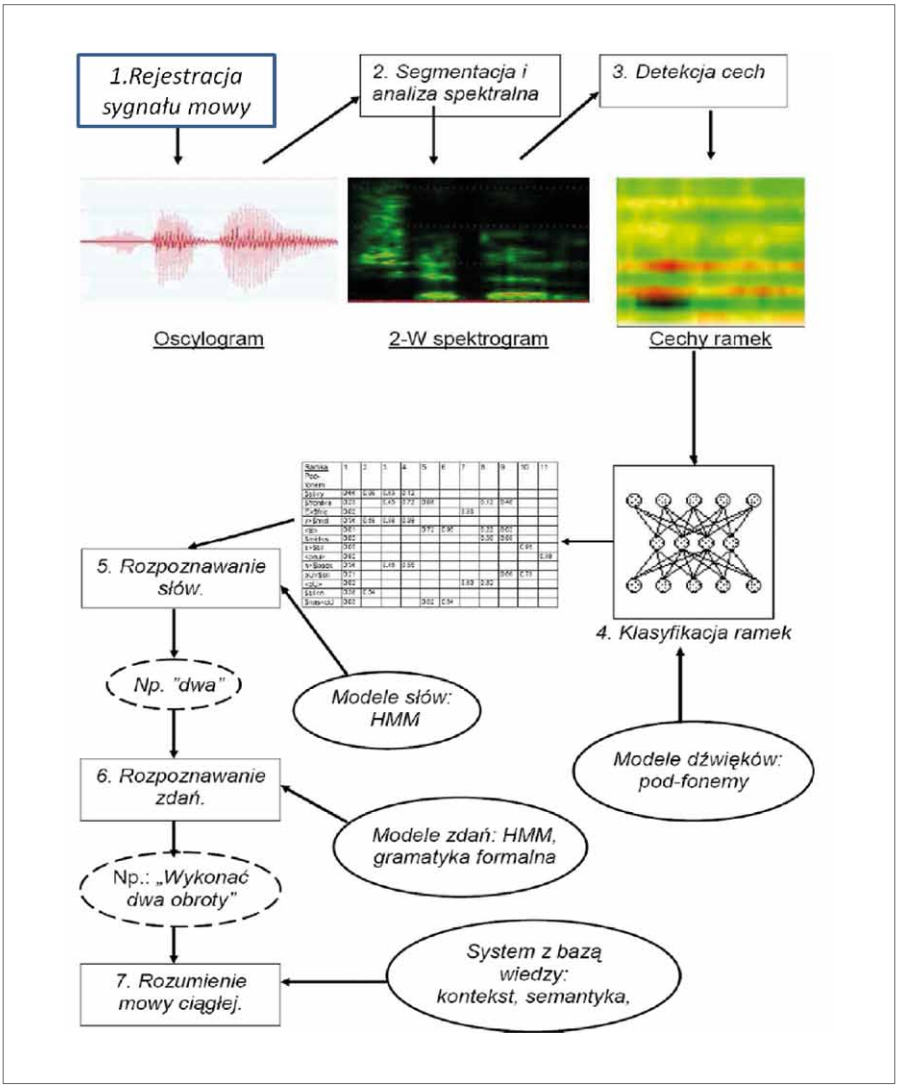
W pierwszej kolejności chodzi o zasoby słownika i o akceptowane reguły gramatyczne. Jedno i drugie jest w takich uproszczonych systemach bardzo limitowane. Na przykład słownik

sterowanego głosem telefonu komórkowego zawiera zaledwie od kilkudziesięciu do kilkuset słów i zwrotów, z reguły z góry narzuconych tak, żeby można je było łatwo i skutecznie rozpoznawać. Wystarcza to w prostych zadaniach sterowania, ale bardzo ogranicza ogólność zastosowań takiego systemu.

Po drugie, systemy budowane według schematu przedstawionego na rysunku 27 z reguły obsługują tylko konkretnych mówców, do których głosów są one specjalnie trenowane. Jeśli pojawi się ktoś nowy, kto będzie chciał użyć swojego głosu do sterowania systemem, to z dużym prawdopodobieństwem taka próba się nie powiedzie. Można to w pewnych okolicznościach uważać za zaletę („Ale wierny – słucha tylko swoich!"), ale w ogólnym przypadku jest to poważne ograniczenie.

Dlatego myśląc o zastosowaniach głosowego sterowania w systemach inteligentnych budynków, musimy sięgać do rozwiązań znacznie bardziej skomplikowanych, ale wolnych od ograniczeń. Taki system o bogatszych możliwościach rozpoznawania mowy przedstawiono na rysunku 28. Przyjrzyjmy się elementom tego rysunku i omówmy krótko ich znaczenie, bo jest to bardzo pouczający przykład, na którym można się opierać, analizując różne inne (pod względem szczegółów) rozwiązania wprowadzane przez innych badaczy lub producentów tego typu urządzeń. Przy omawianiu skorzystamy z faktu, że podobnie jak i przy poprzednim omawianym systemie – poszczególne bloki na schemacie 28 są ponumerowane – i do tych numerów będziemy teraz nawiązywać.

1. Proces rejestracji sygnału mowy (i ewentualnej filtracji, preemfazy, dekonwolucji itp.) był już wcześniej wzmiankowany i omawiany, przeto tylko odnotowujemy jego konieczną obecność w systemie. W rezultacie otrzymujemy cyfrową reprezentację czasowego przebiegu sygnału mowy, którą na rysunku reprezentuje pokazany oscylogram.
2. Kolejnym elementem jest analiza spektralna, której celem jest uzyskanie dynamicznego widma sygnału (patrz rys. 25), co także było już omawiane. Na rysunku wynik tej operacji określany jest jako 2-W spektrogram, czyli widmo dwuwymiarowe. Natomiast w bloku tym wpisana jest jeszcze jedna czynność, na którą warto zwrócić uwagę: segmentacja. Najogólniej mówiąc, chodzi tu o podzielenie ciągłego sygnału mowy na kawałki, które będą podlegały oddzielnemu rozpoznawaniu, a potem z tych oddzielnych rozpoznań montuje się rozpoznanie całości. Takie postępowanie jest konieczne, bo całej długiej i złożonej wypowiedzi jednorazowo rozpoznać się nie da. Na temat tego, jakie segmenty należy wyróżniać w sygnale mowy podczas jego rozpoznawania, napisano już całe tomy, więc nie będziemy w tym momencie przesądzać, czy wydzielonymi segmentami mają być fonemy (odpowiedniki głosek), sylaby czy może pod-fonemy. Na rysunku użyto określenia „ramka" – i tego będziemy się trzymać, pozostawiając dodefiniowanie ramki twórcom implementującym system w konkretnym zastosowaniu.
3. Proces automatycznego rozpoznawania (czegokolwiek!) odwołuje się zawsze do jakichś cech rozpoznawanych obiektów. Dla sygnału mowy badano setki zestawów różnych cech. Aktualnie w charakterze cech do rozpoznawania mowy najpowszechniej używane są tak zwane parametry mel-cepstralne, ale nie ma możliwości żeby tu wyjaśnić, co ten termin oznacza.
4. Mając sygnał mowy podzielony na rozpoznawane ramki oraz mając wybrane cechy, za pomocą których ramki te są przedstawiane w procesie rozpoznawania, możemy przeprowadzić (dla każdej ramki osobno!) klasyfikację, czyli przypisanie do tej ramki identyfikatora, który ją będzie dalej reprezentował. W przypadku, jeśli rozpoznawanymi ramkami są fonemy




**Rys. 28.** Budowa bardziej uniwersalnego systemu rozpoznawania mowy

(lub ich części, tak zwane pod-fonemy), jako identyfikatory rozpoznanych fragmentów mogą być wykorzystane symbole, jakie tym fonemom nadaje się podczas tzw. transkrypcji fonematycznej wypowiedzi w danym języku. Przy innej definicji ramki identyfikatory muszą być wymyślone *ad hoc*. Na rysunku 28 w bloku dokonującym klasyfikacji narysowano schemat sieci neuronowej, co sugeruje, że właśnie to użyteczne narzędzie może być użyte do klasyfikacji, ale możliwe jest także użycie innych klasyfikatorów, których w sztucznej inteligencji wymyślono bardzo wiele.

5. Mając sygnał mowy przekształcony do postaci sekwencji identyfikatorów ramek (wraz z prawdopodobieństwami przypisanymi różnym możliwym rozpoznaniom – patrz tabelka wychodząca z bloku nr 4), możemy próbować z tych pozbawionych jeszcze sensu elementów montować zrozumiałe rozpoznania elementów mowy. W bloku nr 5 jest to montowanie ramek w pojedyncze słowa – najczęściej w oparciu o model HMM.
6. Kolejny blok montuje ze słów kompletne zdania (być może zrozumiałe komendy).
7. Ostatni blok aspiruje do tego, by rozumieć mowę ciągłą – o czym także będzie dalej mowa.

Warto zwrócić uwagę, że na omawianym schemacie kilkakrotnie pojawiło się słowo „model". W nowoczesnych systemach rozpoznawania mowy model jest najważniejszy. Używa się wielu modeli. Mogą to być modele dźwięków, pokazujące, jakie konfiguracje cech ramek pozwalają je rozpoznać jako określone obiekty fonetyczne, na przykład pod-fonemy, istnieją modele HMM, dzięki którym rozpoznajemy słowa, składając je ze zidentyfikowanych ramek, są modele zdań, w których określoną rolę odgrywa gramatyka, jest wreszcie model wiedzy, pozwalający rozumieć sens dłuższych wypowiedzi prezentowanych w formie mowy ciągłej. Zagadnienia te są jednak na tyle specjalistycznie związane z zaawansowanymi technikami rozpoznawania mowy, że nie mieszczą się w ramach tego przeglądowego referatu, w związku z czym zainteresowanych Czytelników musimy odesłać do obszernej i łatwo dostępnej literatury, jaka istnieje na ten temat.

## 6. Próba podsumowania

W referacie wskazano na korzyści, jakie można osiągnąć, stosując w inteligentnych budynkach systemy komunikacji z użytkownikami bazujące na automatycznym rozpoznawaniu mowy. Ponieważ technologia mowy i języka nie jest jeszcze szeroko znana i powszechnie stosowana, w referacie omówiono możliwości i ograniczenia współczesnych systemów automatycznego rozpoznawania mowy oraz skrótowo pokazano ich budowę. Podsumowując te rozważania, warto jeszcze wskazać na trudności, jakie występują przy próbach przenoszenia na grunt polski systemów istniejących dla innych języków (angielskiego, niemieckiego, japońskiego). Niestety każdy język naturalny ma swoją daleko posunięta specyfikę i rozwiązania dobrze sprawdzające się przy automatycznym rozpoznawaniu jednego języka mogą całkowicie zawodzić przy próbie zastosowania do innego języka. Dlatego praktyka, która funkcjonuje dosyć powszechnie w elektronice, automatyce i informatyce, polegająca na adaptowaniu do polskich warunków innowacyjnych rozwiązań opartych na badaniach naukowych prowadzonych za granicą – tu

nie da się zastosować. Autor tego referatu w monografii wydanej w 1978 roku [4] pisał: „Jeśli systemy automatycznego rozpoznawania mowy polskiej mają znaleźć zastosowania w polskiej technice i w polskiej gospodarce – to muszą powstać w oparciu o badania naukowe prowadzone tu, nad Wisłą". Niestety trzeba przyznać, że mimo wielu lat intensywnych badań wciąż jeszcze systemy rozpoznawania mowy polskiej pozostawiają wiele do życzenia, zwłaszcza jeśli system ma akceptować głosy wielu użytkowników, a także wtedy, gdy zamiast rozpoznawania pojedynczych komend przewiduje się konieczność rozpoznawania kierowanych do systemu wypowiedzi formułowanych w postaci sekwencji mowy ciągłej. Nie wchodząc w szczegóły, można stwierdzić, że istnieje pełna analogia pomiędzy zadaniami automatycznego rozpoznawania mowy i zadaniami automatycznego rozpoznawania pisma. W jednym i drugim przypadku rozpoznanie izolowanych elementów (pojedynczych słów lub oddzielnie starannie pisanych znaków) jest relatywnie łatwe, co jednak nie przekłada się wcale na osiągalność rozwiązań globalnych, pozwalających rozpoznawać mowę ciągłą lub ciągłe odręczne pismo (także bardzo niestaranne, jak przysłowiowe zapiski lekarzy na receptach).

Na nasze szczęście (jako osób korzystających na co dzień z komunikacji głosowej) nasz własny system rozpoznawania mowy działa w sposób absolutnie genialny. Bez najmniejszego trudu rozpoznajemy to, co mówi do nas inny człowiek, nawet wtedy, gdy celowej komunikacji głosowej towarzyszą różne zakłócenia i zniekształcenia. Ten fakt utrudnia jednak ludziom zrozumienie trudności, z jakimi styka się konstruktor systemu przeznaczonego do automatycznego rozpoznawania mowy dla potrzeb komunikacji człowieka z maszyną. O tym, że trudno jest komputerowo rozpoznać niestarannie napisany tekst, nikogo przekonywać nie trzeba. Nie trzeba też specjalnie agitować ludzi, żeby starali się pisać wyraźnie, gdy wypełniony przez nich formularz ma być poddawany automatycznemu rozpoznawaniu przez automatyczny czytnik. Natomiast niesłychanie trudno jest zapanować nad skłonnością ludzi do niestarannej i nadmiernie szybkiej wypowiedzi w sytuacji, kiedy mówią do maszyny zdolnej do rozpoznawania mowy. Zwykle wygląda to tak, że użytkownik najpierw nieśmiało rzuca głosem jakieś polecenie i z przyjemnym zdziwieniem odkrywa, że system poprawnie je zinterpretował. Potem śmielej (i coraz mniej starannie!) podawane są kolejne komendy. Wreszcie użytkownik zaczyna mówić swobodnie całymi zdaniami – i w tym momencie komunikacja się urywa, bo system automatyczny osiąga kres swoich możliwości.

Zatem jest jeszcze wiele do zrobienia, jeśli chcemy dysponować w przyszłości inteligentnymi budynkami, którym będzie można wydawać polecenia w języku polskim. Jednak omówione w referacie zalety takiego rozwiązania są na tyle istotne, że warto pracować nad systemami automatycznego rozpoznawania mowy jako nad modułami wygodnie komunikującymi ludzi z systemami automatyki. W tym obszarze czeka nas jeszcze długa droga i konieczne jest pokonanie wielu trudności, ale podjąć odpowiednie wysiłki zdecydowanie warto.

Przygotowując prezentowany tu artykuł, zawężono tematykę, pomijając i eliminując z pola widzenia dodatkowo także wszystkie te metody i techniki akustyczne, fonetyczne i informatyczne, które związane są ze sztuczną generacją sygnału mowy. Wprawdzie w skład każdej komunikacji głosowej wchodzi zarówno odbiór mowy rozmówcy, jak i głosowe odpowiedzi, a ponadto sztuczna synteza mowy ma generalnie wiele zastosowań, w tym także może być wykorzystana w automatyce, jednak ten wątek w niniejszej pracy całkowicie eliminujemy, odsyłając zainteresowanego Czytelnika do pozycji [5] bibliografii.


## Podziękowanie

Praca niniejsza powstała w ramach programu Badań Statutowych Katedry Automatyki AGH, umowa numer 11.11.120.612.


## Literatura


[1] Nejat Ince A. (Editor): *DIGITAL SPEECH PROCESSING Speech Coding, Synthesis and Recognition*. Kluwer Academic Publishers, 1992.

[2] Laface P., De Mori R.: *Speech Recognition and Understanding*. Springer-Verlag, Berlin – Heidelberg 1992.

[3] Tadeusiewicz R.: *Sygnał mowy*. WKiŁ, Warszawa 1988 (Monografia książkowa dostępna obecnie także w Internecie: http://winntbg.bg.agh.edu.pl/skrypty/0004/).

[4] Tadeusiewicz R.: *Głosowa łączność człowieka z maszyną cyfrową*. ZN AGH, Seria Monografie, Automatyka nr 22, Kraków 1978.

[5] Kondoz A.M.: *Digital Speech Coding for Low Bit Rate Communications Systems*. John Wiley & Sons Ltd, 1994.

[6] Deller J.R., Proakis J.G., Hansen J.H.L.: *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.

[7] Keller E.: *Fundamentals of Speech Synthesis and Speech Recognition*. John Willey & Sons Ltd., 1994.

[8] Gold B., Morgan N.: *Speech and Audio Signal Processing*. John Wiley & Sons, Ltd., 2000.

[9] Quatieri T.F.: *Discrete-Time Speech Signal Processing*. Prentice Hall, 2002.

[10] Chu W.C.: *Speech Coding Algorithms Foundation and Evolution of Standardized Coders*. John Wiley & Sons Ltd., 2003.

[11] Markowitz J.A.: *Using Speech Recognition*. Prentice Hall, 1996.

[12] Rabiner R., Juang B.H.: *Fundamentals of Speech Recognition*. Prentice Hall, 1993.




**Ryszard Tadeusiewicz** – AGH – Akademia Górniczo-Hutnicza, Katedra Automatyki i Inżynierii Biomedycznej

artykuł recenzowany