Karolina Kuligowska[1], Paweł Kisielewicz[2],
Aleksandra Włodarz[3]

Wady i ograniczenia systemów rozpoznawania mowy

1. Wstęp

Rozpoznawanie mowy (ang. *speech recognition*, *speech-to-text* – STT) jest procesem umożliwiającym przekształcenie wypowiedzianych słów i zdań na tekst w postaci cyfrowej. Zaprojektowanie maszyny, która naśladuje zdolność człowieka do słuchania, fascynowało badaczy od stuleci. Jednakże dopiero w ciągu ostatniej dekady dziedzina technologii rozpoznawania mowy dynamicznie się rozwinęła, a obecnie budowane zaawansowane systemy potrafią reagować na spontaniczną mowę w języku naturalnym.

Celem niniejszego artykułu jest zbadanie głównych ograniczeń wpływających na dokładność rozpoznawania mowy. W rozdziale drugim przedstawiono pokrótce klasyfikację systemów rozpoznawania mowy. Rozdział trzeci zawiera opis działania istniejących systemów rozpoznawania mowy. W rozdziale czwartym scharakteryzowano stosowane podejścia i techniki wykorzystywane w rozpoznawaniu mowy. W rozdziale piątym przeanalizowano wady i ograniczenia systemów rozpoznawania mowy. Rozdział szósty stanowi podsumowanie artykułu.

2. Klasyfikacja systemów rozpoznawania mowy

Najbardziej intuicyjne rozróżnienie między systemami rozpoznawania mowy dotyczy rodzaju rozpoznawanej mowy, sposobu obsługi mówcy oraz wielkości

[1] Uniwersytet Warszawski, Wydział Nauk Ekonomicznych.
[2] Profeosoft, Kraków.
[3] Profeosoft, Kraków.

słownika[4]. Biorąc pod uwagę rodzaj rozpoznawanej mowy, wyróżnia się: rozpoznawanie izolowanych słów (ang. *isolated words*), rozpoznawanie słów łączonych (ang. *connected words*), rozpoznawanie mowy ciągłej (ang. *continuous speech*) i rozpoznawanie mowy spontanicznej (ang. *spontaneous speech*).

W przypadku sposobu obsługi mówcy (ang. *speaker dependence*) systemy dzielą się na te, które potrafią rozpoznać tylko konkretnego mówcę (ang. *speaker dependent system*), te, które potrafią rozpoznać dowolnego mówcę (ang. *speaker independent system*) oraz wreszcie na te, które adaptują się do konkretnego mówcy (ang. *speaker adaptable system*)[5].

Słownik jest zbiorem wyrazów, które mają być rozpoznane, dlatego tak ważny jest jego rozmiar. Jeżeli liczba słów w słowniku jest bardzo mała, ale słowa te znacząco różnią się między sobą pod względem akustycznym, system może osiągnąć bardzo wysoki poziom dokładności rozpoznawania mowy. Im większy słownik, tym więcej niejasności wynikających z liczby możliwych alternatywnych sposobów wymowy danego słowa. Małe słowniki mogą zwierać poniżej 30 słów, z kolei większość dużych systemów rozpoznawania mowy zawiera słowniki o rozmiarach kilku tysięcy słów. Słowniki systemów przeznaczonych do dyktowania i transkrypcji mogą zawierać 10.000 słów i więcej. Nawet tak duży rozmiar słownika może nie być wystarczający, gdyż w świecie rzeczywistym nigdy nie da się przewidzieć, jakie słowa wypowie użytkownik[6].

3. Działanie systemów rozpoznawania mowy

Tworzone obecnie systemy rozpoznawania mowy opierają swoją architekturę na podobnych modułach składowych, które przetwarzają dane wejściowe według określonych reguł. W pierwszym kroku sygnał mowy zostaje poddany wstępnej obróbce. Ze względu na to, iż dane dźwiękowe obarczone są nadmiarem informacji w postaci szumu powstałego podczas pobierania próbki oraz

[4] Vrinda, Ch. Shekhar, *Speech recognition system for English language*, „International Journal of Advanced Research in Computer and Communication Engineering" 2013, vol. 2, issue 1, s. 919.

[5] Ibidem, s. 920.

[6] R.E. Gruhn, W. Minker, S. Nakamura, *Statistical Pronunciation Modeling for Non-Native Speech Processing*, *Signals and Communication Technology*, Springer-Verlag, Berlin Heidelberg 2011, s. 13.

przydźwięku sieci, faza ta redukuje ilość zbędnych informacji. Tak przygotowane dane wejściowe poddawane są dalszej analizie.

Kluczowe fazy w rozpoznawaniu mowy to segmentacja i parametryzacja. Przed rozpoczęciem procesu rozpoznawania, sygnał mowy powinien zostać podzielony na małe segmenty, czyli na słowa, fonemy (najczęściej stosowany podział) lub sylaby, ale może być też segmentowany na fragmenty zależnie od cech fonetycznych. Ważne przy segmentacji mowy jest również oddzielenie mowy od ciszy[7]. Spośród wielu metod najczęściej stosowana jest segmentacja równomierna, która dzieli sygnał na równe fragmenty o z góry określonych długościach[8]. Jej zaletą jest łatwość implementacji, jednak nie bierze ona pod uwagę tego, gdzie zaczyna się, a gdzie kończy fonem. Dlatego aby uzyskać dokładniejsze modelowanie, bardziej użyteczna może okazać się segmentacja nierównomierna, która lokalizuje granice fonemów.

Gdy sygnał mowy jest już podzielony, w celu rozróżnienia dźwięków wyodrębnienia się z sygnału mowy istotne informacje przez parametryzację. Opiera się ona na ekstrakcji cech charakterystycznych tego sygnału (ang. *feature extraction*). Oznacza to konwersję sygnału mowy na wektory cech takich jak na przykład amplituda, moc, intensywność, podstawowa częstotliwość, wykorzystywane do dalszej analizy i przetwarzania. Najważniejsze techniki stosowane w ekstrakcji cech to: analiza cepstralna – *Cepstral Analysis* (CA), analiza mel-cepstralna – *Mel Cepstrum Analysis* (MCA), melowo-częstotliwościowe współczynniki cepstralne – *Mel-frequency cepstral coefficients* (MFCC), liniowa analiza dyskryminacyjna – *Linear Discriminant Analysis* (LDA), liniowe kodowanie predykcyjne – *Linear Predictive Coding Analysis* (LPC), percepcyjna predykacja liniowa – *Perceptual Linear Predictive Analysis* (PLP)[9].

Dekoder, czyli silnik odpowiedzialny za odnalezienie najlepszego dopasowania w bazie wiedzy na podstawie przychodzących wektorów cech, wykonuje rzeczywistą decyzję w rozpoznaniu wypowiedzi, łącząc i optymalizując informacje przekazywane przez modele akustyczne i językowe. Silniki rozpoznawania mowy dopasowują wykryte słowo do słowa już znanego ze zgromadzonej bazy wiedzy za pomocą jednej z następujących technik[10]:

---

[7] B. Ziółko, M. Ziółko, *Przetwarzanie mowy*, Wydawnictwa AGH, Kraków 2011, s. 224–225.

[8] M. Tarasiuk, Z. Gosiewski, *Segmentacja mowy polskiej z wykorzystaniem transformacji falkowej*, „Acta Mechanica et Automatica" 2010, vol. 4, no. 1, s. 92.

[9] N. Desai, K. Dhameliya, V. Desai, *Feature Extraction and Classification Techniques for Speech Recognition: A Review*, „International Journal of Emerging Technology and Advanced Engineering" 2013, vol. 3, issue 12, s. 368.

[10] S. Gaikwad, B. Gawali, P. Yannawar, *A review on Speech Recognition Technique*, „International Journal of Computer Applications" 2010, vol. 10, no. 3, s. 22.

- Dopasowywanie całych słów (ang. *whole-word matching*)
  Silnik porównuje przychodzący sygnał audio-cyfrowy z wcześniej zapisanym szablonem słowa. Ta technika wymaga, by użytkownik (lub ktoś inny) nagrał wcześniej te słowa, które będą rozpoznawane. Czasami oznacza to nagranie nawet kilkuset tysięcy słów. Wzory całych słów wymagają dużej ilości pamięci (pomiędzy 50 i 512 bajtów na słowo) i sprawdzają się tylko wtedy, gdy słownik rozpoznania jest znany już podczas rozwijania aplikacji.
- Dopasowywanie podsłów (ang. *sub-word matching*)
  Silnik szuka podsłów, zwykle fonemów, a następnie wykonuje na nich dalsze rozpoznawanie wzorców. Technika ta wymaga więcej przetwarzania niż technika dopasowywania całych słów, ale wymaga dużo mniej pamięci (pomiędzy 5 i 20 bajtów na słowo). Ponadto wymowa danego słowa może zostać odgadnięta z tekstu danego języka, bez potrzeby wcześniejszego nagrywania słów.

4. Podejścia i techniki wykorzystywane w rozpoznawaniu mowy

W technologii rozpoznawania mowy można wyróżnić trzy główne kategorie podejść do rozpoznawania, a mianowicie[11]:

1. Podejście akustyczno-fonetyczne (ang. *acoustic-phonetic approach*), które zakłada, że jednostki fonetyczne są charakteryzowane przez szereg cech, takich jak na przykład częstotliwość, tonacja, barwa dźwięku. Cechy te są wydobywane z sygnału mowy i wykorzystywane między innymi przy segmentacji mowy. Podejście to jest rzadko stosowane w aplikacjach komercyjnych.
2. Podejście wykorzystujące rozpoznawanie wzorców (ang. *pattern recognition approach*), które obejmuje dwa niezbędne etapy: trenowanie wzorców (ang. *pattern training*) oraz porównanie wzorców (ang. *pattern comparison*). Ważną cechą takiego podejścia jest to, że wykorzystuje ono dobrze sformułowany aparat matematyczny oraz ustanawia spójne reprezentacje wzorców mowy dla wiarygodnego porównywania wzorców, od zestawu oznakowanych próbek treningowych po formalny algorytm treningowy.
3. Podejście wykorzystujące wiedzę (ang. *knowledge based approach*), nazywane również podejściem bazującym na sztucznej inteligencji (ang. *Artificial*

[11] T. Shanthi, L. Chelpa, *Review of Feature Extraction Techniques in Automatic Speech Recognition*, „International Journal of Scientific Engineering and Technology" 2013, vol. 2, issue 6, s. 483.

Intelligence approach). Polega ono na zmechanizowaniu procedury rozpoznania mowy w sposób zbliżony do tego, jak dokonuje tego człowiek, wykorzystując posiadaną wiedzę dotyczącą między innymi cech akustycznych. Podejście to jest połączeniem podejścia akustyczno-fonetycznego i podejścia wykorzystującego rozpoznawanie wzorców.

5. Wady i ograniczenia systemów rozpoznawania mowy

Na dokładność systemów rozpoznawania mowy wpływa wiele czynników. Poniżej przeanalizowano te z nich, które mają istotny wpływ na dokładność rozpoznania mowy.

5.1. Środowisko rozpoznawania mowy

Wydajność rozpoznawania mowy drastycznie spada w hałaśliwym otoczeniu, gdyż zakłócenia w przestrzeni powodują rozbieżności pomiędzy warunkami treningowymi (czystymi) oraz warunkami, w jakich odbywa się rozpoznawanie (hałaśliwymi). Zakłócenia zniekształcają i zanieczyszczają sygnał mowy oraz zmieniają wektory danych reprezentujących mowę. Badania prowadzone nad problemem odporności na zakłócenia skupiają się na dwóch kierunkach[12]:

a) usuwanie szumu z zakłóconego hałasem sygnału poprzez filtr szumów, odejmowanie widmowe, filtr Wienera, filtr RASTA, mapowanie wektorów stochastycznych;
b) kompensowanie efektu szumu w przestrzeni modelu akustycznego, dopasowując środowisko treningowe do warunków realizacyjnych, w jakich odbywa się rozpoznawanie mowy poprzez mapowanie wektorów stochastycznych oraz sekwencyjne szacowanie hałasu.

5.2. Urządzenie rejestrujące głos

Gdy mikrofon nie jest wystarczająco czuły albo zbyt wrażliwy, może wygenerować informację audio, która będzie trudna do rozszyfrowania. Do izolowania głosów od szumów często stosuje się zestaw mikrofonów, gdzie czysty sygnał

[12] C.-H. Wu, C.-H. Liu, *Robust Speech Recognition for Adverse Environments*, w: *Modern Speech Recognition Approaches with Case Studies*, S. Ramakrishnan (red.), Intech 2012, s. 3.

mowy przechwycony przez kilka mikrofonów jest oddzielony od hałaśliwego sygnału. Zestaw mikrofonów można kierować w najbardziej dogodną stronę, co poprawia wydajność rozpoznawania, jednak zakłada to synchroniczną i ciągłą obserwację sygnału[13], co nie zawsze jest możliwe do uzyskania.

5.3. Prozodia

Prozodia jest istotna w zrozumieniu języka mówionego: ułatwia rozpoznać wypowiadane słowa, globalne i lokalne dwuznaczności oraz analizować strukturę dyskursu. Jednakże cechy prozodyczne nie są wykorzystywane w większości współczesnych systemów rozpoznawania mowy[14]. Informacje prozodyczne są trudne do modelowania i nadal szuka się rozwiązań, które pomogłyby przezwyciężyć problem prozodii w kontekście systemów automatycznego rozpoznawania mowy.

5.4. Zmienność mowy

Każdy człowiek ma swój indywidualny sposób mówienia, inny ton i barwę głosu, mówi w innym tempie oraz rytmie, inaczej artykułuje wyrazy, używa innego języka. Zmienność mowy determinują także wszelkie wady wymowy i problemy z dykcją, różnice demograficzne, kulturowe i geograficzne[15] oraz wiek, przynależność klasowa i akcent.

5.5. Styl mowy

Systemy rozpoznawania izolowanych słów wymagają krótkich pauz pomiędzy wypowiadanymi słowami. Taki styl mówienia nie jest naturalny, dlatego systemy te tracą na swej popularności na rzecz systemów rozpoznawania mowy ciągłej, na których obecnie skupiona jest większość badań[16,17]. W spontanicznej,

---

[13] K. Machida, A. Ito, *Speech recognition under noisy environments using multiple microphones based on asynchronous and intermittent measurements*, 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2013), IEEE, 2013, s. 163.

[14] L. Mary, *Extraction and Representation of Prosody for Speaker, Speech and Language Recognition*, SpringerBriefs in Electrical and Computer Engineering, Springer, 2012, s. 17–19.

[15] F. Biadsy, *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*, rozprawa doktorska, Department of Computer Science, Columbia University, 2011, s. 1–2.

[16] R. Buttermore, N. Lee-Perkins, *Developments in voice recognition technology*, 13th Annual Freshman Engineering Conference, 2013, s. 3–4.

[17] S. Renals, T. Hain, *Speech recognition*, w: *The Handbook of Computional Linguistics and Natural Language Processing*, A. Clark, Ch. Fox, S. Lappin (red.), Wiley-Blackwell, 2010, s. 300–301.

swobodnej wypowiedzi lub pod presją czasu bardzo często dochodzi do redukcji wymowy niektórych fonemów lub sylab, co może doprowadzić do utraty części informacji i niesie za sobą wyższy wskaźnik błędów podczas rozpoznawania.

5.6. Mowa dzieci

Dzieci, w porównaniu z dorosłymi, mają krótszą krtań oraz fałdy głosowe. Skutkuje to słabą rozdzielczością widmową dźwięków głosu oraz nieliniowym wzrostem formantowych częstotliwości. Sporym problemem jest też nieprawidłowa wymowa dzieci. Bardzo często nie znają one poprawnych form fleksyjnych określonych słów, szczególnie tych, które są wyjątkami do ogólnie przyjętych zasad. Pomimo tego, iż zaproponowano kilka technik, które miały za zadanie poprawę dokładności systemów rozpoznawania w przypadku dziecięcych głosów, wydajność takich systemów jest dużo niższa niż w przypadku rozpoznawania mowy dorosłego człowieka[18].

5.7. Ograniczony słownik

Dużo pracy wymaga stworzenie i rozwinięcie dobrego słownika. Większość systemów automatycznego rozpoznawania mowy działa z dużym, lecz zwykle ograniczonym słownikiem, znajdującym najlepiej pasujące słowa dla danego sygnału akustycznego. Podczas gdy systemy rozpoznawania mowy ciągłej tworzą wysokiej jakości transkrypcję, nie radzą sobie z rozpoznawaniem słów spoza słownika. Jeśli słowa w języku wykazują zmienność morfologiczną, konieczne może okazać się rozszerzenie słownika nawet do setek tysięcy słów. Jednak niektóre języki posiadają tak bogatą morfologię, że modelowanie języka wymagałoby słownika, który wykracza rozmiarem ponad rozsądną wielkość. W takich przypadkach, aby skonstruować słownik, najlepiej zrezygnować z modelowania opartego na słowach, a wykorzystać podsłowa, na przykład morfemy.

5.8. Kontekst i homonimy

Jednym z problemów rozpoznawania mowy jest określenie kontekstu, w jakim słowa zostały wymówione. Niektóre słowa, które brzmią bardzo podobnie, mogą zostać dobrze rozpoznane tylko wtedy, gdy znany jest ich kontekst. Dodatkowo

[18] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont i in., *Automatic speech recognition and speech variability: A review*, „Speech Communication" 2007, no. 49, s. 767–768.

kontekst wpływa na dokładność rozpoznania homonimów – wyrazów o takim samym brzmieniu, lecz różnym znaczeniu. Systemy rozpoznawania mowy nie mają możliwości odróżnienia ich na podstawie samego dźwięku. Określenie kontekstu danego słowa wpływa pozytywnie na dokładność rozpoznania i wydajność systemów pod względem rozróżniania homonimów.

5.9. Różnorodność języków

Różnorodność języków stawia wyzwania przed rozpoznawaniem mowy. Języki aglutynacyjne mają bogatszy zasób słownictwa (a tym samym większe słowniki) ze względu na tworzenie słów, polegające na łączeniu ze sobą wielu morfemów. Z kolei języki fleksyjne charakteryzują się stosunkowo swobodnym szykiem zdania oraz bardzo bogatym systemem morfologicznym i derywacyjnym. Dokładna analiza cech dystynktywnych danego języka ułatwia wybór metody rozpoznawania mowy. Rodzaj fonemów występujących w danym języku, warianty alofoniczne, wzory sylab oraz cechy fleksyjne decydują o tym, jaką technikę zastosować dla rozpoznawania mowy w danym języku[19].

5.10. Języki słowiańskie

Większość systemów rozpoznawania mowy operuje na najbardziej rozpowszechnionych w świecie językach, takich jak angielski, francuski, niemiecki czy japoński. Języki słowiańskie w dalszym ciągu czekają na intensywny rozwój technologii mowy pod ich kątem. Jednym z wyzwań jest fleksyjna natura języków słowiańskich, która modyfikuje podstawową formę elementów leksykalnych zgodnie z relacjami gramatycznymi, morfologicznymi i kontekstowymi. Liczba powstawania wielu form wyrazowych bardzo często przekracza milion odrębnych pozycji, które muszą zostać uwzględnione i właściwie zarządzane. Różnica ta jest bardzo duża w porównaniu z systemami rozpoznawania mowy zaprojektowanymi na przykład dla języka angielskiego, w którym słownik 50 tys. najczęściej używanych słów daje wskaźnik pokrycia 99%. Języki słowiańskie wymagają na ogół słowników, które są od 10 do 20 razy większe[20].

---

[19] S. Saraswathi, T.V. Geetha, *Design of language models at various phases of Tamil speech recognition system*, „International Journal of Engineering, Science and Technology" 2010, vol. 2, no. 5, s. 244–245.

[20] J. Nouza, J. Zdansky, P. Cerva, J. Silovsky, *Challenges in Speech Processing of Slavic Languages (Case Studies in Speech Recognition of Czech and Slovak)*, w: *Development of Multimodal*

6. Podsumowanie i kierunki dalszych badań

Rozpoznawanie mowy jest nowoczesną technologią, w rozwijaniu której badacze borykają się z pewnymi ograniczeniami we współczesnych podejściach. Spośród wad i ograniczeń rozpoznawania mowy najwięcej uwagi w najnowszej literaturze poświęca się dokładności. Dlatego też w niniejszym artykule przeanalizowano dziesięć najistotniejszych czynników, które wpływają na dokładność rozpoznawania mowy, a mianowicie: środowisko rozpoznawania mowy, urządzenie rejestrujące głos, prozodia, zmienność mowy, styl mowy, mowa dzieci, ograniczony słownik, kontekst i homonimy, różnorodność języków, języki słowiańskie.

Poza dokładnością rozpoznawania mowy, w dalszych badaniach należy się przyjrzeć zjawiskom przepustowości i opóźnienia. Poprawa przepustowości jest bardzo istotna w implementacji aplikacji o dużych wymaganiach czasowych, takich jak aplikacje multimedialne. Natomiast poprawa opóźnienia pozwala systemom rozpoznającym mowę na sprawne działanie w czasie rzeczywistym.

Bibliografia

Benzeghiba M., De Mori R., Deroo O., Dupont S. i in., *Automatic speech recognition and speech variability: A review*, „Speech Communication" 2007, no. 49, s. 767–768.

Biadsy F., *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*, rozprawa doktorska, Department of Computer Science, Columbia University, 2011, s. 1–2.

Buttermore R., Lee-Perkins N., *Developments in voice recognition technology*, 13th Annual Freshman Engineering Conference, 2013, s. 3–4.

Desai N., Dhameliya K., Desai V., *Feature Extraction and Classification Techniques for Speech Recognition: A Review*, „International Journal of Emerging Technology and Advanced Engineering" 2013, vol. 3, issue 12, s. 368.

Gaikwad S., Gawali B., Yannawar P., *A review on Speech Recognition Technique*, „International Journal of Computer Applications" 2010, vol. 10, no. 3, s. 22.

Gruhn R.E., Minker W., Nakamura S., *Statistical Pronunciation Modeling for Non-Native Speech Processing*, *Signals and Communication Technology*, Springer-Verlag, Berlin Heidelberg 2011, s. 13.

Interfaces: Active Listening and Synchrony, A. Esposito, N. Campbell, C. Vogel i in., (red.), Springer-Verlag, Berlin Heidelberg 2010, s. 227–228.

Machida K., Ito A., *Speech recognition under noisy environments using multiple microphones based on asynchronous and intermittent measurements*, Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013, s. 2.

Mary L., *Extraction and Representation of Prosody for Speaker, Speech and Language Recognition*, SpringerBriefs in Electrical and Computer Engineering, Springer, 2012, s. 17–19.

Nouza J., Zdansky J., Cerva P., Silovsky J., *Challenges in Speech Processing of Slavic Languages (Case Studies in Speech Recognition of Czech and Slovak)*, w: *Development of Multimodal Interfaces: Active Listening and Synchrony*, A. Esposito, N. Campbell, C. Vogel i in. (red.), Springer-Verlag, Berlin Heidelberg 2010, s. 227–229.

Renals S., Hain T., *Speech recognition*, w: *The Handbook of Computional Linguistics and Natural Language Processing*, A. Clark, Ch. Fox, S. Lappin (red.), Wiley-Blackwell, 2010, s. 299–301.

Saraswathi S., Geetha T.V., *Design of language models at various phases of Tamil speech recognition system*, „International Journal of Engineering, Science and Technology" 2010, vol. 2, no. 5, s. 244–245.

Shanthi T., Chelpa L., *Review of Feature Extraction Techniques in Automatic Speech Recognition*, „International Journal of Scientific Engineering and Technology" 2013, vol. 2, issue 6, s. 483.

Tarasiuk M., Gosiewski Z., *Segmentacja mowy polskiej z wykorzystaniem transformacji falkowej*, „Acta Mechanica et Automatica" 2010, vol. 4, no. 1, s. 92.

Vrinda, Shekhar Ch., *Speech recognition system for English language*, „International Journal of Advanced Research in Computer and Communication Engineering" 2013, vol. 2, issue 1, s. 919.

Wu C.-H., Liu C.-H., *Robust Speech Recognition for Adverse Environments*, w: *Modern Speech Recognition Approaches with Case Studies*, S. Ramakrishnan (red.), Intech 2012, s. 3.

Ziółko B., Ziółko M., *Przetwarzanie mowy*, Wydawnictwa AGH, Kraków 2011, s. 224–225.

* * *

Speech Recognition Systems: Disadvantages and Limitations

Summary

Speech is one of the easiest ways allowing communication between people and machines. Speech recognition technology makes everyday life easier: it is widely used in mobile phones, computers, tablets, cars, etc. However, the quality of automatic speech recognition is affected by many factors, therefore, so much effort is put into improving the performance of speech recognition systems. The aim of this paper is to present the current state of development of speech recognition systems and to examine

their drawbacks and limitations. The paper discusses the current classification, construction and functioning of speech recognition systems, which gives an insight into the speech-to-text software implemented so far. The analysis of disadvantages and limitations of speech recognition systems has allowed identifying the weak points of these systems. Problems that are to be solved in the near future indicate the direction of further development of speech recognition systems.

Keywords: speech recognition system, speech-to-text, STT, speech recognition limitations.