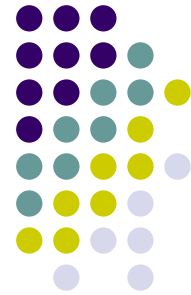


Transformacje głosu

Akustyka mowy



Modyfikacja parametrów sygnału mowy prowadząca do zmiany charakteru głosu mówcy przy zachowaniu treści wypowiedzi.



Zastosowania

- Syntezatory mowy
- Sztuka filmowa
- Gry komputerowe
- Śpiew
- Tłumaczenia na inne języki i nauka języków obcych
- Anonimizacja mówcy
- Zmiana przekazywanych emocji
- „Oszukiwanie” systemów rozpoznawania mówcy



Cechy dystyngtywne mówcy

W ujęciu socjo/psychologicznym: styl mówienia, emocje

- intonacja
- akcent
- wzorce długości poszczególnych dźwięków i pauz
- głośność i tempo wypowiedzi
- sposób artykulacji

Prozodia silnie związana jest z treścią wypowiedzi, ale można zdefiniować wzorce, które charakteryzować będą konkretnego mówcę:

- pochylenie frazy
- pochylenie początkowe
- pochylenie końcowe
- średnia wartość pochylenia akcentu
- parametry czasowe, określające rytm wypowiedzi (wzorce długości poszczególnych dźwięków i pauz)

Cechy dystyngtywne mowy



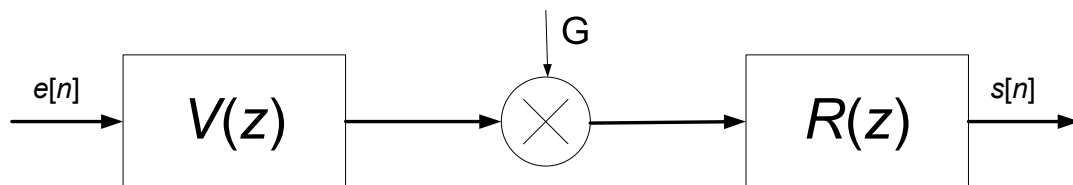
W ujęciu fizjologicznym: indywidualne cechy głosu

- parametry tonu krtaniowego
 - wartość średnia F_0 (wyznaczana dla określonej bazy nagrań głosu mówcy) – „naturalna wysokość głosu”
 - zakres częstotliwości, w ramach którego zmienia się częstotliwość podstawowa (min-max, trzykrotna wartość odchylenia standardowego od wartości średniej)
 - kształt impulsów tonu krtaniowego
- parametry traktu głosowego
 - częstotliwości środkowe, szerokości pasm i wzmocnienie formantów
 - wypowiedanie konkretnych fonemów wpływa głównie na częstotliwości środkowe formantów $F1$ i $F2$
 - formant $F3$ i wyższe zależą głównie od długości traktu głosowego (cecha indywidualna mówcy) i ich częstotliwości środkowe zmieniają się niewiele podczas wypowiedania różnych głosek
 - pasma wszystkich formantów zależą głównie od cech osobniczych

Modelowanie sygnału mowy



Liniowy model pobudzenie-filtr



$e[n]$ – pobudzenie

$s[n]$ – sygnał mowy

$V(z)$ – transmitancja traktu głosowego

$R(z)$ – charakterystyka promieniowania ust

G – wzmocnienie

Model sinusoidalny



pobudzenie: $e(t) = \sum_{k=1}^N a_k(t) \cos \varphi_k(t)$ $\varphi_k(t) = \int_0^t \omega_k(\tau) d\tau + \phi_k$

trakt głosowy: $H(\Omega, t) = M(\Omega, t) \exp[j\Psi(\Omega, t)]$
 $M_k(t) = M[\omega_k(t), t]$ $\Psi_k(t) = \Psi[\omega_k(t), t]$

sygnał mowy: $s(t) = \sum_{k=1}^N A_k(t) \cos[\theta_k(t)]$
 $A_k(t) = a_k(t) M_k(t)$ $\theta_k(t) = \varphi_k(t) + \Psi_k(t)$

Możliwe modyfikacje



- Modyfikacje tempa wypowiedzi
- Modyfikacje częstotliwości podstawowej
- Modyfikacje traktu głosowego
- Konwersja głosu
- Morphing głosu

Modyfikacja tempa wypowiedzi



Skalowanie czasu – zmiana tempa wypowiedzi przy zachowaniu charakterystyki widmowej sygnału

Funkcja skalująca: $\beta(t)$

Modyfikacja: $D(t) = \int_0^t \beta(\tau) d\tau$

Zakładając: $t_a^{i+1} = t_a^i + P(t_a^i)$

Mapowanie: $t_a^i \leftrightarrow t_s^i$

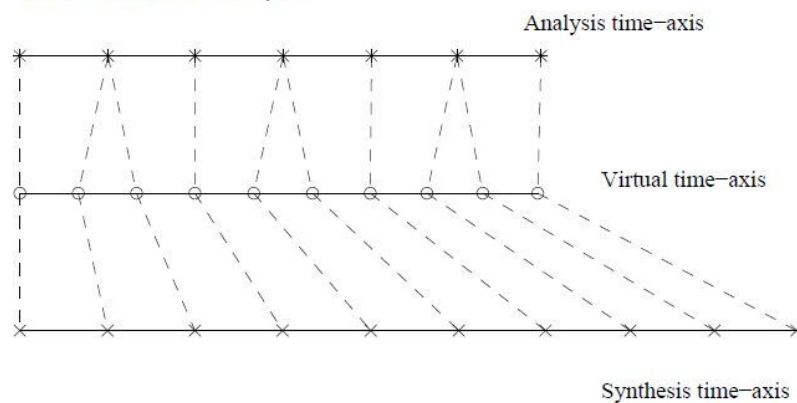
$$t_s^{i+1} - t_s^i = \frac{1}{t_u^{i+1} - t_u^i} \int_{t_u^i}^{t_u^{i+1}} P(t) dt$$

gdzie: $t_s^i = D(t_u^i)$

Modyfikacja tempa wypowiedzi



Time-scale modification by 1.5



Modyfikacja częstotliwości podstawowej



Skalowanie częstotliwości – zmiana charakterystyki widmowej sygnału przy zachowaniu tempa wypowiedzi

Oryginalny kontur częstotliwości podstawowej (okresu):

$$P(t)$$

Funkcja skalująca:

$$\alpha(t)$$

Zakładając:

$$t_a^{i+1} = t_a^i + P(t_a^i)$$

Mapowanie:

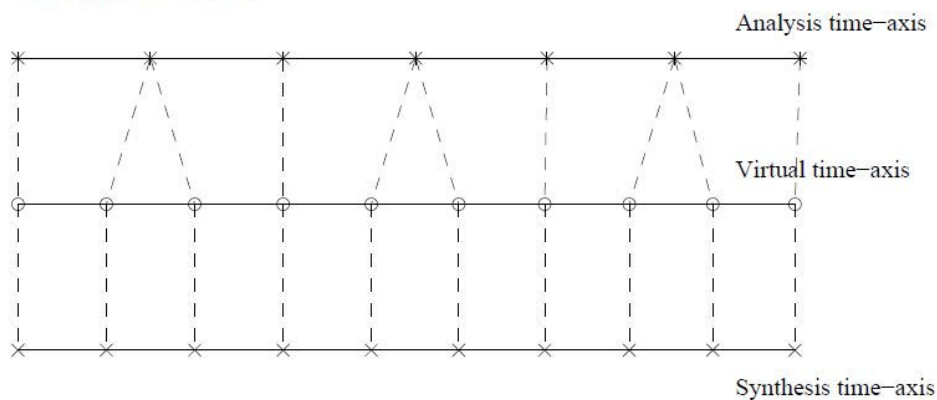
$$t_a^i \leftrightarrow t_s^i$$

$$t_s^{i+1} - t_s^i = \frac{1}{t_s^{i+1} - t_s^i} \int_{t_s^i}^{t_s^{i+1}} \frac{P(t)}{\alpha(t)} dt$$

Modyfikacja częstotliwości podstawowej



Pitch modification by 1.5



Modyfikacje częstotliwości podstawowej i tempa wypowiedzi



- Algorytmy działające w dziedzinie czasu i częstotliwości
- Najczęściej stosowane – algorytmy PSOLA (Pitch Synchronous OverLap-Add)
 - TD-PSOLA – w dziedzinie czasu
 - FD-PSOLA – w dziedzinie częstotliwości
- Podobne algorytmy:
 - SOLA
 - WSOLA
 - MBROLA

Skalowanie czasu i częstotliwości PSOLA



- Przetwarzanie sygnału mowy w krótkich segmentach, a następnie odpowiednie ich połączenie
- Aby uniknąć nieciągłości w miejscach łączenia segmentów stosuje się nakładkowanie oraz odpowiednie okna
- Dla sygnałów okresowych (lub prawie okresowych) sensownym jest dopasowanie długości okna do długości okresu
- W algorytmie PSOLA (opracowanym specjalnie dla przetwarzania mowy), długość kolejnych okien dobierana jest zgodnie z wartością estymowanej częstotliwości podstawowej
- Najlepszymi znacznikami początków ramek byłyby chwile zamknięcia głośni, jednak ze względu na trudność ich wyznaczenia stosuje się inne znaczniki (np. CoG)

Skalowanie czasu i częstotliwości

TD-PSOLA



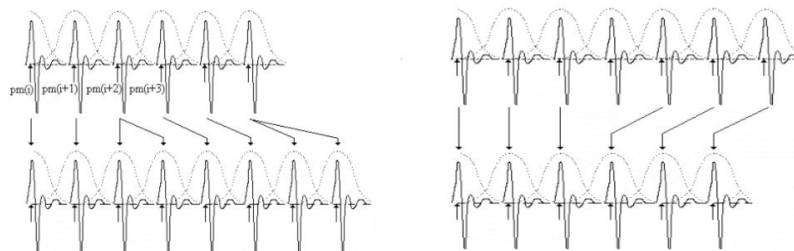
- Estymacja częstotliwości podstawowej
- Podział sygnału mowy na segmenty synchronicznie z estymowaną częstotliwością podstawową (dla bezdźwięcznych fragmentów mowy długość segmentów jest z góry określona i stała).
- Modyfikacja sygnału.
- Rekonstrukcja sygnału poprzez złożenie segmentów z zastosowaniem zakładek.

Skalowanie czasu i częstotliwości

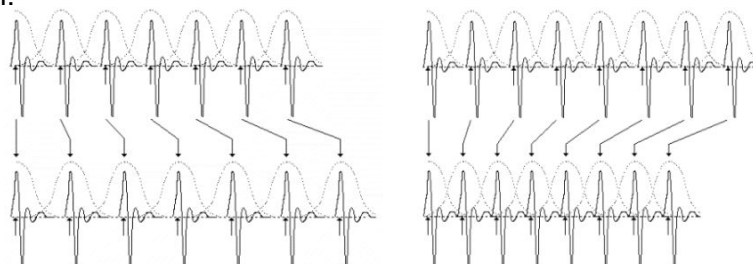
TD-PSOLA



Skalowanie czasu:



Skalowanie częstotliwości:



Skalowanie częstotliwości

FD-PSOLA



- Podział sygnału mowy na segmenty synchronicznie z częstotliwością tonu krtaniowego (nie jest konieczne wyznaczenie dokładnych chwil początków okresów, wystarczy dobry algorytm estymacji częstotliwości podstawowej).
- Obliczenie krótkookresowego widma sygnału, estymowanie obwiedni widma, obliczenie pobudzenia (w tym kroku istotnym jest, by widmo uzyskanego pobudzenia było płaskie, należy więc skorzystać z dobrego algorytmu wyznaczania obwiedni widma).
- Modyfikacje częstotliwości podstawowej.
- Rekonstrukcja sygnału (przejście z dziedziny częstotliwości do dziedziny czasu - może się okazać, że po dokonaniu transformacji nie istnieje sygnał rzeczywisty, który odpowiadałby uzyskanemu widmu - należy wyznaczyć widmo, dla którego istnieje sygnał rzeczywisty, a które jest jak najbardziej zbliżone do widma syntetycznego).

Skalowanie częstotliwości

FD-PSOLA



Sposoby modyfikacji częstotliwości:

- usuwanie (dla obniżenia) lub dodawanie (dla podwyższenia) harmonicznym w widmie sygnału - konieczna jest dokładna estymacja częstotliwości podstawowej dla wyznaczenia harmonicznym, muszą być zachowane zależności fazowe między poszczególnymi harmonicznymi;
- kompresja/ekspansja widma pobudzenia - oryginalna oś częstotliwości jest „zawijana” (ang. *warping*) z wykorzystaniem współczynnika skalującego β (wprowadza zniekształcenia, jednak jeśli obwiednia widma została wyznaczona poprawnie, będą one mniejsze niż w przypadku usuwania/dodawania harmonicznym)

$$Y(k_s) = (1 - \alpha)X(k_v) + \alpha X(k_v + 1)$$

$$\alpha = k_s - \frac{k}{\beta}$$

Skalowanie czasu i częstotliwości

wykorzystanie modelu sinusoidalnego



Skalowanie czasu: $t' = \beta t$

Skalowanie częstotliwości: $\omega'_k = \beta \omega_k$

Częstotliwość można również przeskalować skalując najpierw czas, a następnie przepróbkując uzyskany sygnał (mniejsza złożoność obliczeniowa).

Dyskusja



Algorytmy TD:

- Szybkie, wymagają małych mocy obliczeniowych – sprawdzają się w systemach czasu rzeczywistego
- Bardzo dobre rezultaty przy małych współczynnikach skalowania
- Problem powtarzania transjentów

Algorytmy FD

- Bardziej złożone obliczeniowo
- Najczęściej wymagają obliczenia parametrów modelu (wyższa jakość)
- Przewyższają algorytmy TD w przypadku dużych współczynników skalowania.

Modyfikacja charakterystyki traktu głosowego



Skorzystanie z algorytmu PSOLA – przepróbkowanie segmentów mowy przed ich ponownym złożeniem

- Aby uzyskać podniesienie częstotliwości środkowych formantów (przeskalowanie przez $\gamma > 1$) należy zmniejszyć częstotliwość próbkowania γ razy.
- Segmenty są dodawane z oryginalną częstotliwością zmienia się więc położenie formantów, ale częstotliwość podstawowa i tempo wypowiedzi pozostają niezmienione.
- Mała złożoność obliczeniowa, ale częstotliwości środkowe formantów można zmieniać tylko liniowo.

Modyfikacja charakterystyki traktu głosowego



- Estymacja charakterystyki traktu głosowego (obwiedni widma sygnału mowy)
- Zamodelowanie charakterystyki traktu głosowego
- Modyfikacja zgodnie z założonymi regułami
- Synteza

Modyfikacja charakterystyki traktu głosowego



Modyfikacja obwiedni widma sygnału mowy

- Różne techniki estymacji obwiedni widma
- Aby podnieść częstotliwości środkowe formantów γ -krotnie należy zmodyfikować widmo sygnału zgodnie ze wzorem

$$Y(t, \Omega_k) = X(t, \Omega_k) E(t, \Omega_k / \gamma) / E(t, \Omega_k)$$

- Możliwość transformacji nieliniowych (czynnik γ zmienny w czasie)

Modyfikacja charakterystyki traktu głosowego



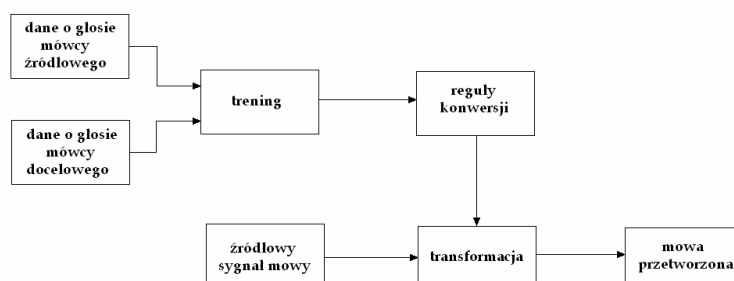
Modyfikacja biegunów filtru modelującego trakt głosowy

- Transmitancja filtru estymowana jest za pomocą predykcji liniowej.
- Bieguny transmitancji zapisywane są w postaci biegunowej $re^{j\varphi}$
- Zmiana kąta φ powoduje przesunięcie formantu na osi częstotliwości.
- Zmiana promienia r powoduje zwężenie lub poszerzenie zajmowanego przez formant pasma.
- Dokonując transformacji należy pamiętać, by bieguny znajdowały się zawsze wewnątrz okręgu jednostkowego.

Konwersja głosu



Automatyczna transformacja głosu mówcy źródłowego do głosu mówcy docelowego z zachowaniem treści wypowiedzi.



System konwersji gromadzi dane o głosach mówcy źródłowego i docelowego (odpowiednie próbki głosów) i na ich podstawie automatycznie generuje reguły konwersji w procesie treningu. Reguły te są następnie wykorzystywane w procesie transformacji głosu źródłowego tak, by odpowiadał charakterystyce głosu docelowego.

Konwersja głosu



Istnieje wiele systemów konwersji mowy, opierających się na różnych modelach mowy i metodach modyfikacji, jednak w każdym podejściu należy rozwiązać trzy podstawowe problemy:

1. wyodrębnienie cech charakterystycznych mówców z przebiegów akustycznych mowy,
2. opracowanie metody mapowania cech mówców źródłowego i docelowego,
3. modyfikacja charakterystyki głosu mówcy źródłowego, tak by brzmiał jak głos mówcy źródłowego, z wykorzystaniem opracowanego wcześniej schematu mapowania.

Konwersja głosu



- Przed analizą z reguły należy również zgromadzić odpowiednią bazę danych wypowiedzi mówcy źródłowego i docelowego (najczęściej te same wypowiedzi) oraz odnaleźć odpowiadające sobie ramki sygnału w wypowiedziach (za pomocą ukrytych modeli Markova lub nieliniowej transformacji czasu).
- Dla mapowania charakterystyk mówców, czyli znajdowania funkcji zależności między cechami mówcy źródłowego i docelowego, większość systemów konwersji korzysta z trzech podstawowych narzędzi: kwantyzacji wektorowej VQ (ang. *Vector Quantization*), liniowej kombinacji rozkładów normalnych GMM (ang. *Gaussian Mixture Model*) i sztucznych sieci neuronowych ANN (ang. *Artificial Neural Networks*).

Morphing głosu



- Analogia do morphingu obrazów.
- Stopniowe przechodzenie od głosu mówcy źródłowego do głosu mówcy docelowego .
- Potrzebne są takie same zdania wypowiedziane przez obu mówców.
- Konieczne jest odnalezienie odpowiadających sobie segmentów w wypowiedziach obu mówców (fonosegmentacja, nieliniowa transformacja czasu DWT).
- Oddzielnie przeprowadzana jest modyfikacja pobudzenia (np. za pomocą technik SOLA) i charakterystyki traktu głosowego (transformata Fouriera, współczynniki LPC, PARCOR...)

Bibliografia



- [1] O. Türk, *New Methods for Voice Conversion*, M.S. Thesis, Electrical and Electronics Engineering, Boğaziçi University, 2000
- [2] Min Tang, Chao Wang and S. Seneff, *Voice Transformations: From Speech Synthesis to Mammalian Vocalizations*, Proc. of the 7th European Conference on Speech Communication and Technology, pp. 353-356, September 2001
- [3] T. F. Quatieri and R. J. McAulay, *Shape Invariant Time-Scale and Pitch Modification of Speech*, IEEE Trans. on Signal Processing, vol. 40, pp. 497 – 510, March 1992
- [4] D. Rentzos, S. Vaseghi, Qin Yan, Ching-Hsiang Ho, *Voice Conversion Through Transformation of Spectral and Intonation Features*, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 1-21-24, May 2004
- [5] J. Slifka and T. R. Anderson, *Speaker Modification with LPC Pole Analysis*, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 644-647, May 1995
- [6] E. Moulines and W. Verhelst, *Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech*, chapter 15 in W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*, Elsevier, 1995
- [7] M. Kahrs and K. Brandenburg, *Applications of Digital Signal Processing to Audio and Acoustics*, Kluwer Academic Publishers, 2002
- [8] J. R. Deller and Proakis and J.H. Hansen, *Discrete-time processing of speech signals*, Macmillan Publishing Company, 1993