خوارزمية البحث عن مسار نبرة الصوت في الإشارة الكلامية

الدكتور المهندس حسان محمد احمد

أستاذ مساعد، كلية هندسة الحاسوب والمعلوماتية، الجامعة السورية الخاصة

الملخص

يُقدم هذا العمل خوارزمية العثور على مسار نبرة الصوت (pitch track) بناءً على خوارزمية مختلطة للبحث في المجالات الطيفية والزمنية (spectral & time domains) للإشارة الأصلية ولتحويلها غير الخطي (nonlinear transformation).

تتشكل مجموعة القيم المرشحة (candidates) لتكوين مسار النبرة عند خرج كلٍ من دالة الارتباط التوافقي الطيفي (spectral harmonic correlation) ودالة الارتباط المتبادل المستنظمة (normalized cross-correlation) للإشارة الأصلية.

بعد الفرز النهائي للقيم المرشحة في المجموعة الحاصلة، يتم تشكيل المسار النهائي لنبرة الصوت الموافقة للإشارة الكلامية.

الكلمات المفتاحية: الإشارة الكلامية، نبرة الصوت، دالة الارتباط.

Pitch track search algorithm in speech signal

Abstract

This article describes an algorithm for search a pitch track based on a mixed search algorithm in the spectral and time domains for the original signal and its nonlinear transformation. The set of candidates is formed at the output of the spectral harmonic correlation function and the normalized cross-correlation function. After the final screening of candidates, the final track is formed.

Keywords: speech signal, pitch, correlation function,

1. مقدمة

يعدُّ التردد اللحظي للنبرة الأساسية (F_0) البارامتر الأهم في تصنيف الإشارات الكلامية في التمثيل البارامتري للكلام (speech parametric representation)، والذي يُعرّف على أنّه التردد اللحظي لتذبذبات الحبال الصوتية (vocal cords) للمتكلم. من أهم المؤشرات الرئيسية لجودة التقييم، والتي تتمثل في دقة الوقت والتردد، أي سرعة الاستجابة لتغيرات F_0 ومقدار الانحراف، والتي تقوم بتحديدها الخوارزميات المستخدمة في معالجة الإشارة الكلامية [5،6].

إن تباين تردد النبرة الأساسية كبير جدًا، ويمكن أن يختلف اختلافًا كبيرًا ليس فقط بين الأشخاص (لأصوات الذكور يكون التردد $70-200\,\mathrm{Hz}$ ، ولأصوات الإناث يمكن أن يصل إلى $400\,\mathrm{Hz}$)، وإنما أيضًا لشخص واحد وخاصة في الكلام الإنفعالي [1،19].

حتى الآن، هناك العديد من الخوارزميات المقترحة لتقييم النبرة الأساسية للصوت، أي حدة ارتفاع الصوت، بما في ذلك تلك التي تستخدم طرق التقييم في كلٍ من مجالات الزمن والتردد [9،10،11،16،28]، وأكثر هذه الخوارزميات شيوعًا هي RAPT [24] و YIN [2،8] و SWIPE [7] وتعديلاتها. وعلى الرغم من معدل الخطأ المنخفض، حتى في وجود الضجيج (سواءً الخلفي أو الشرطي الناتج عن الإثارة المختلطة للقناة الصوتية) ، فإنّ دقة التقييم تنخفض مع تعديل F_0 (modulation).

تُعدّ طريقة الارتباط التلقائي (autocorrelation) الطريقة الرئيسية التي ظهرت لآحقًا على أساسها عائلة كاملة من خوارزميات تحديد نبرة الصوت في الإشارة الكلامية. والنهج بسيط للغاية – من الضروري حساب دالة الارتباط التلقائي وأخذ أول قيمة عُظمى لها، والتي ستعرض المكون الترددي الأكثر وضوحًا في الإشارة [13،14] .

2. هدف البحث

نظرًا لأن مسألة تحديد تردد النبرة الأساسية (F_0) تبرز تقريبًا، بشكل أو بآخر، أمام كل شخص يعمل مع الصوت والكلام، فلابد من إيجاد طريقة أكثر فعالية من الطرق المتاحة لحلها.

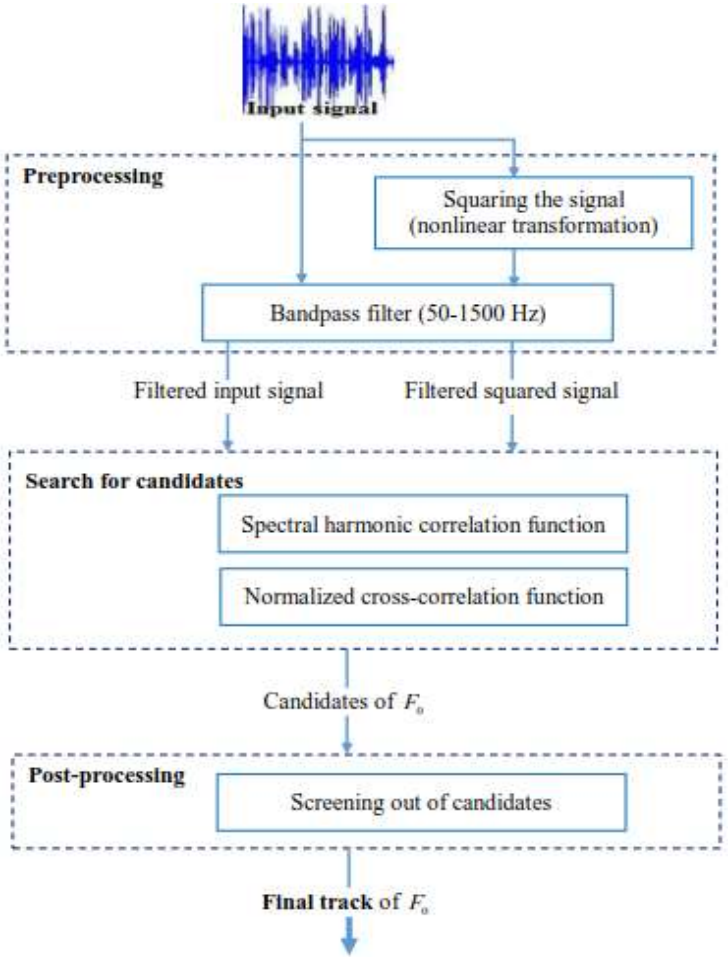
يتجلى الهدف الرئيس من هذا العمل في إيجاد خوارزمية تقوم بتقليل الحساسية تجاه تعديلات التردد الأساسي لنبرة الصوت ودرجة الضجيج في الإشارة الكلامية من خلال دمج طريقة الارتباط (correlation method) واختيار التردد (frequency selection) لتقييم التردد اللحظي للنبرة الأساسية (F_0)، وتعمل على تحديد مسار نبرة الصوت في الإشارة الكلامية.

3. مواد وطرق البحث

بعد أن وضعنا أمامنا مسألة تقليل الحساسية تجاه تعديلات التردد الأساسي لنبرة الصوت ودرجة الضجيج في الإشارة الكلامية، فإن طريقتنا المقترحة لتشكيل مسار النبرة تستخدم دالة الارتباط المتبادل المستنظمة (normalized cross-correlation function, NCCF) في البحث عن القيم المرشحة لتشكيل مسار النبرة وتركيبة من طريقة الارتباط واختيار التردد F_0. لذلك، ولأجل تحقيق الاستقرار ضد التشويشات الخارجية، نقوم بإجراء التقييم في المنطقة الطيفية للإشارة الأصلية ولتحويلها غير الخطي.

3.1. توصيف الخوارزمية المقترحة

يمكن تحديد الخطوات الرئيسية للخوارزمية وفق الآتي: المعالجة المسبقة (Preprocessing) للإشارة، والبحث عن القيم المرشحة (Search for candidates) لتشكيل مسار النبرة، والمعالجة النهائية اللاحقة (final post-processing). يبين الشكل (1) المخطط العام للخوارزمية المقترحة.

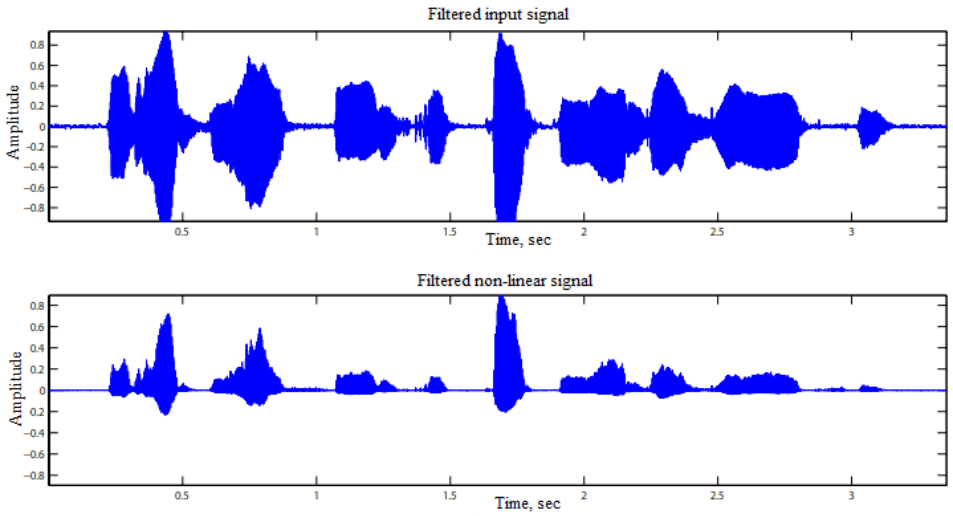


الشكل (1): المخطط العام لخوارزمية تحديد مسار النبرة

3.1.1. المعالجة المسبقة

يظهر التردد الأساسي F_0 عند استخدام مربّع الإشارة حتى لو كان المطال (amplitude) صغيرًا أو غائبًا في البيانات الأولية، كما هو موضح في [26]، وهذا مميز للكلام في الخطوط الهاتفية. وبالتالي، فإن المعالجة المسبقة تشمل إنشاء نسخة من الإشارة الأصلية وتحويلها غير الخطي، أي التربيع (squaring) والاستنظام

(normalization)، وكذلك الترشيح اللاحق للإشارات الأصلية والمربعة باستخدام مرشح تمرير حزمة (band pass filter, BPF) بعرض حزمة ترددات (bandwidth) من ($50-1500\,\text{Hz}$) [15،18]. يتم تحديد الفاصل الزمني (interval) على التردد F_0 ضمن المجال ($60-400\,\text{Hz}$). يبين الشكل (2) نتائج المعالجة المسبقة للإشارة الكلامية.



الشكل (2): الإشارة الأصلية واللاخطية بعد الترشيح

3.1.2. البحث عن القيم المرشحة لـ F_0 وفقًا للقيمة العظمى لدالة الارتباط التوافقي الطيفي، SHC.

يعتمد أساس طريقة اختيار التردد على افتراض أنه أثناء الإثارة الصوتية للقناة الصوتية (vocal tract)، يحتوي طيف الإشارة على ذروات (peaks) عند الترددات التي تعد مضاعفات تردد النبرة الأساسية (pitch frequency) [27]. يتم إجراء البحث على فترات زمنية ($32\,\text{ms}$) مع تراكب (overlap) بمقدار ($10\,\text{ms}$) وتردد أخذ عينات (sampling frequency) ($16\,\text{kHz}$).

للحصول على تبيين تردد (frequency resolution) أفضل، يتم تطبيق عملية الاستيفاء (interpolation) باستخدام مرشح الكتروني مثالي نافذي (windowed sinc filter)، الذي يقوم بحذف كل الترددات في طيف الإشارة والتي هي أعلى من تردد القطع الأساسي (cutoff frequency) ويبقي حزمة التردد المنخفضة للإشارة، ونحصل بذلك على خطوة التردد (frequency step) من ($7.8\,\mathrm{Hz}$) وعرض النافذة من 2048 عينة (samples) [24]. بعد ذلك، يتم إنشاء دالة الارتباط التوافقي الطيفي (spectral harmonic correlation function, SHC)، والتي تحددها العلاقة التالية:

$$SHC(n, f) = \sum_{-W/2}^{W/2} \prod_{r=1}^{R} S(n, r\ f + W/2) \qquad (1)$$

حيث:

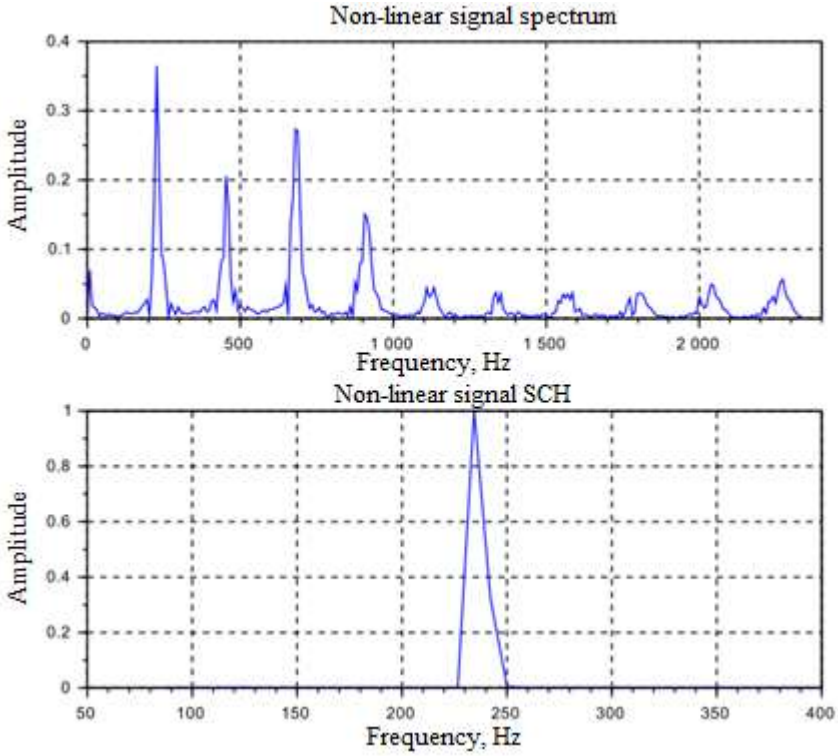
$S(n, f)$ – طيف الإشارة للإطار (frame) n؛ f – تردد الإشارة (frequency)؛

W – عرض النافذة (window width)؛ R – عدد التوافقيات (number of harmonics).

نظرًا لأن الإشارة مستنظمة، فإن القيمة القصوى للدالة SHC هي 1. يتم إجراء البحث عن القيمة العظمى المحلية فقط من أجل طيف الإشارة المربعة، وتحدد قيمة العتبة (0.6) لاستبعاد القيم العظمى الخاطئة. يبين الشكل (3) طيف ودالة الارتباط المتبادل الطيفي لإطار من الإشارة.

لتقليل الأخطاء ، يتم حساب F_0 في المقاطع الصوتية. لأجل تحديد نوع الفاصل الزمني، يتم استخدام نسبة الطاقة الترددية المنخفضة المستنظمة (normalized low–frequency energy ratio, NLFER) [20،21]، والتي يتم تحديدها من خلال نسبة مجموع المكونات الطيفية للإطار في النطاق الترددي $F_{0\max} - F_{0\min}$ إلى متوسط القيمة على الإشارة بأكملها، وفق العلاقة التالية:

$$NLFER(n) = \frac{\sum_{f=F_{0\min}}^{F_{0\max}} S(n,f)}{\frac{1}{N}\sum_{n=1}^{N}\sum_{f=F_{0\min}}^{F_{0\max}} S(n,f)} \qquad (2)$$



الشكل (3): طيف ودالة الارتباط المتبادل الطيفي لإطار من الإشارة

3.1.3. البحث عن القيم المرشحة لـ F_0 وفقًا للقيمة العظمى لـ NCCF

يتم حساب القيم المرشحة لتحديد مسار النبرة لكلٍ من الإشارة الأصلية $(s(n))$ والإشارة المعدلة بشكل غير خطي $(s(n+k))$ باستخدام دالة الارتباط المتبادل المستنظمة (NCCF) وفق العلاقة التالية:

$$NCCF(k) = \frac{1}{\sqrt{e_0 e_k}} \sum_{n=1}^{N-K_{\max}} s(n)s(n+k) \qquad (3)$$

حيث:

$$e_0 = \sum_{n=1}^{N-K_{\max}} s(n)^2, \qquad e_k = \sum_{n=k}^{k+N-K_{\max}} s(n)^2, \quad K_{\min} \le k \le K_{\max}$$

تتوافق القيم القصوى المحلية لـ NCCF مع تأخر (delay) الإشارة الذي يساوي دور (period) النبرة الأساسية. في حالة وجود بعض القيم القصوى المحلية لـ NKKF القريبة بالقيمة من 1، يتم تحديد القيمة الموافقة لأصغر دور. نظرًا لأن القيم الموجودة في المقاطع غير الصوتية أقل بكثير من 1، يتم حساب NKKF فقط في المقاطع الصوتية وفقًا للعلاقة (3).

3.1.4. المعالجة اللاحقة النهائية

في مرحلة المعالجة اللاحقة النهائية، يتم البحث عن دائرة (contour) قيم النبرة الأساسية باستخدام البرمجة الديناميكية (dynamic programming)، التي تربط القيم المرشحة لدور النبرة التي تم العثور عليها في المجالين الطيفي والديناميكي، وبذلك يُفرض قيد على أن تردد النغمة الأساسية يتغير ببطء، وبالتالي، يجب ألا تختلف قيم ترددات الإطارات المتجاورة اختلافًا كبيرًا.

4. نتائج التجارب

أجريت جميع التجارب والاختبارات في بيئة ماتلاب (Matlab) وباستخدام صندوق الأدوات البرمجية المفتوحة المصدر [25,31].

4.1. قاعدة البيانات الكلامية

من المهم جدًا أن يتم اختبار خوارزميات تحديد تردد نبرة الصوت الأساسية (F_0) ومسارها في الإشارة الكلامية على ذات قواعد البيانات الكلامية. يتوفر العديد من قواعد

البيانات الكلامية مفتوحة المصدر والتي تم إنشاؤها من قبل بعض مخابر الأبحاث المتخصصة، ومنها:

- "The Pitch-Tracking Database": تشمل 2342 عبارة تم تسجيلها بأصوات 10 رجال و 10 نساء [17]؛
- " The fundamental frequency determination algorithm evaluation database": تشمل 50 عبارة تم تسجيلها مرة بصوت رجل، ومرة أخرى بصوت إمرأة [4].
- " LibriSpeech ASR corpus ": تشمل مايقارب 1000 ساعة قراءة باللغة الإنكليزية بتردد ($16\,\text{kHz}$) [29].
- "RAVDESS ": تشمل 1440 ملفًا صوتياً بتردد ($48\,\text{kHz}$) [30].

تتضمن البيانات تسجيلات صوتية (تم تسجيلها باستخدام ميكروفونات خاصة (laryngophone)) وقيم لترددات معيارية مرجعية لنبرة الصوت الأساسية تم حسابها وفقًا للمسارات (trajectory) من الميكريفون الخاص.

تم في هذا العمل استخدام قاعدة البيانات " The Pitch-Tracking Database"، واحتساب الترددات المرجعية في نافذة بعرض ($32\,\text{ms}$) وبتراكبات من ($10\,\text{ms}$).

4.2. مسار تردد النبرة الأساسية

يبين الشكل (4) نتيجة اختبار الخوارزمية المقترحة والممثلة بالمسار النهائي لنبرة الصوت في الإشارة الكلامية المستخدمة.

يتمثل مقياس الأخطاء في نسبة الأخطاء الجسيمة (Gross Error, GE) التي يمكن حسابها بالعلاقة التالية:

$$GE = \frac{1}{N_{VF}} \sum_{k=1}^{N_{VF}} \delta\left(F_0^{ref}(t), F_0^{est}(t)\right) \qquad (4)$$

حيث:

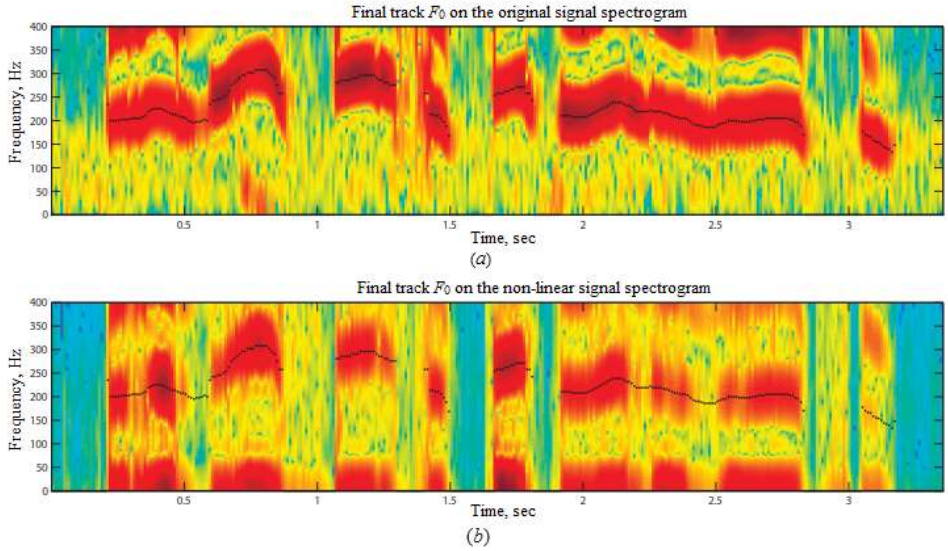
$$\delta\left(F_0^{ref}(t), F_0^{est}(t)\right) = \begin{cases} 1, & \left|\frac{F_0^{ref}(t) - F_0^{est}(t)}{F_0^{ref}(t)}\right| > 0.2 \\ 0 & \end{cases}$$

N_{VF} – عدد الإطارات الصوتية (vocalized frames)؛

F_0^{ref} – القيمة المرجعية (reference value) للتردد النبرة الأساسية F_0 ؛

F_0^{est} – القيمة التقديرية (estimated value) للتردد النبرة الأساسية F_0 ؛

بالتالي، يتم تحديد الإطارات مع التقييم الحاصل بأكثر من 20% .



الشكل (4): (*a*) – طيف الإشارة الأصلية والمسار النهائي للتردد F_0

(*b*) – طيف التحويل اللاخطي للإشارة والمسار النهائي للتردد F_0

5. الاستنتاجات والتوصيات

نظرًا لأن مسألة تحديد تردد النبرة الأساسية (F_0) تبرز بشكل أو بآخر تقريبًا أمام كل شخص يعمل مع الصوت والكلام، فإننا نجد العديد من الطرق لحلها. يتم تحديد مسألة الدقة المطلوبة وخصائص المواد الصوتية المستخدمة في كل حالة محددة وفقًا

لضرورة تحديد البرامترات بعناية، أو يمكن الاقتصار على حل المسألة باستخدام خوارزمية معروفة.

تم تنفيذ الطريقة المقترحة لإيجاد مسار النبرة الأساسية للصوت في الإشارة الكلامية على أساس خوارزمية البحث المشترك في المجالين الطيفي والزمني للإشارة الكلامية الأصلية ولتحويلها غير الخطي. وقد بين التجارب أن فعالية الطريقة تعود إلى استخدام النسخة اللاخطية للإشارة للبحث عن القيم المرشحة لتحديد مسار نبرة الصوت الأساسية ودمج نتائج البحث.

وبنتيجة الاختبارات، بلغت قيمة الأخطاء الجسيمة $GE = 3.75\%$ للذكور، و $GE = 3.45\%$ للإناث. بالتالي، بلغت دقة تحديد مسار النبرة في الإشارة الكلامية للذكور 96.25% وللإناث 96.55% .

يمكن استخدام الخورازمية المقترحة لتحديد تردد النبرة الأساسية لحل مجموعة واسعة من المسائل:

- التعرف على الحالة الانفعالية للشخص (emotion recognition)؛
- تحديد جنس المتكلم (gender speaker determination)؛
- تجزئة الصوت أو تقسيم الكلام إلى جمل (speech segmentation, speech dividing into phrases)؛
- في الطب، لتحديد الخصائص المرضية للصوت (على سبيل المثال، التعرف على علامات مرض باركنسون [20]).

6. المراجع

1. Akhmad, H M 2007 – **Vvedenie v cifrovuyu obrabotku rechevyh signalov**. Ucheb. posobie / H. M. Ahmad, V. F. Zhirkov; Vladim. gos. un-t. – Vladimir: Izd-vo Vladim. gos. un-ta, – 192 s. – ISBN 5-89368-751-5 [Rus]
2. AULIA, F, BASUKI, A, DEWNTARA, B S B **2020** Implementation of Yin Algorithm to Detect Human Voice Emotions According to Gender: Implementation of Yin Algorithm to Detect Human Voice Emotions According to Gender. **Jurnal Mantik**, 4(1), pp. 709-717.
3. AZAROV, I S, VASHEVICH, M I, PETROVSKIY, A A 2012 Algoritm ocenki mgnovennoj chastoty osnovnogo tona rechevogo signala // **Cifrovaya obrabotka signalov**, № 4. S. 49–57. [Rus]
4. BAGSHAW, P C, MILLER, S M, JACK, M A 1993 Enhanced pitch tracking and the processing of the F0 contours for computer aided intonation teaching // **Proceedings of EUROSPEECH, Berlin, Germany**, 1003–1006. http://www.cstr.ed.ac.uk/research/projects/fda
5. BEAT GFELLER, CHRISTIAN FRANK et al. **2020** SPICE: Self-supervised Pitch Estimation. **IEEE trans. On audio, speech and language processing**, CASSP-2020.
6. BRATA, I P B W & DARMAWAN, I D M B A **2021** Comparative study of pitch detection algorithm to detect traditional Balinese music tones with various raw materials. **J. Phys.: Conf. Ser. 1722 012071**.
7. CAMACHO, A, HARRIS, J G 2008 A saw tooth waveform inspired pitch estimator for speech and music // **Journal Acoust. Soc. Am. 2008**, Vol. 123, № 4. P. 1638–1652.
8. CHEVEIGNE, A, HIDEKI KAWAHARA 2002 YIN, a fundamental frequency estimator for speech and music.

Computer Science, Medicine The Journal of the Acoustical Society of America. J. Acoust. Soc. Am. 111 (4).

9. GERHARD, D 2003 Pitch Extraction and Fundamental Frequency: History and Current Techniques. **Technical report, Dept. of Computer Science, University of Regina-2003**.
10. HERMES, D J 1993 – **Pitch analysis / Visual Representations of Speech Signals** / edited by M. Cooke, S. Beet, M.C. Wiley. 1993. P. 325.
11. HESS, W J 1992 – **Pitch and voicing determination / Advances in Speech Signal Processing** /edited by S. Furui, M.M. Sohndi. 1992. P. 348.
12. KAVITA, K, ZAHORIAN, S 2002 Yet another algorithm for pitch tracking // **Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference**. IEEE 2002. Vol. 1. P. 1—361.
13. KOLOKOVO, A S, Lyubinskij, I A **2019** Izmerenie osnovnogo tona rechevogo signala s ispolzovaniem funkcii avtokorrelyacii. **Avtomat. i telemeh**, , vypusk 2, stranicy 152–160. [Rus]
14. MANPREET KAUR, GAGADEEP KAUR, PRIYNKA SOOD **2020** Analysis of performance of pitch estimation techniques. **Journal of critical reviews**, VOL 7, ISSUE 17.
15. ORCHISAMA DAS, JULIUS, SMITH, O, CHRIS CHAFE **2020** Improved Real-time Monophonic Pitch Tracking with the Extended Complex Kalman Filter. **Journal of the Audio Engineering Society**, Vol 68, No. 1/2.
16. PAVLOVETS, A, PETROVSKY, A 2011 Robust HNR-based closed loop pitch and harmonic parameters estimation // **Proc. the 12th Annual Conference of the International Speech Communication Association (Interspeech-2011)**, Italy, Florence, 27-31 August 2011.
17. PIRKER, G, WOHLMAYR, M, PETRIK, S et al 2012 Database for multi-pitch tracking // Graz University of Technology,

Signal Processing and Speech Communication Laboratory. http://www2.spsc.tugraz.at/databases/PTDB-TUG/

18. PRADEEP RENGASWAMY, KIRAN REDDY et al **2020**. Robust f0 extraction from monophonic signals using adaptive sub-band filtering. **Speech Communication**, Volume 116, January 2020, Pages 77-85.
19. PRAJWAL, S, KHOUSHIKH, S et al **2020** A Comparative Study of Various Pitch Detection Algorithms. **IEEE, 5th International Conference on Computing, Communication and Security (ICCCS).**
20. RUSZ, J, CMEJLA, R, RUZICKOVA, H 2011 Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. **The Journal of the Acoustical Society of America**, vol. 129, issue 1, pp. 350-367.
21. RYHOR VASHKEVICH, ELIAS AZAROV **2020** Pitch-invariant Speech Features Extraction for Voice Activity Detection. **IEEE, 22th International Conference on Digital Signal Processing and its Applications (DSPA)**.
22. STEPHEN, A, ZAHORIAN, PRINCY DIKSHIT, HONGBING HU 2006 A spectral-temporal method for pitch tracking. **INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing**, Pittsburgh, PA, USA, September 17-21.
23. STEVEN, W SMITH 1999 – **Digital Signal Processing**. California Technical Publishing, Second Edition.
24. TALKIN, D 1995 – **A Robust Algorithm for Pitch Tracking (RAPT) / Speech Coding and Synthesis** / W.B. Kleijn, K.K. Paliwal eds. Elsevier, ISBN 0444821694. 1995.
25. MIKE BROOKS, **VOICEBOX: Speech Processing Toolbox for MATLAB**. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

26. ZAHORIAN, S A, HU H 2008 A spectral/temporal method for robust fundamental frequency tracking // **The Journal of the Acoustical Society of America**. № 123. P. 4559–4571.
27. ZHUKOVA, A B, MASLENNIKOV, A L **2019** Voice pitch frequency detection via spectrum peaks search with additional frequency weight functions. **Politechnical student journal**, no. 12(41)
28. ZUBRYCKI, P, PETROVSKY, A 2010 Quasi-periodic signal analysis using harmonic transform with application to voiced speech processing // **ISCAS 2010**: 2374-2377.
29. VASSIL PANAYOTOV et al 2015 LibriSpeech: an ASR corpus based on public domain audio books. **ICASSP 2015**. http://www.openslr.org/12 ;
30. LIVINGSTONE SR, RUSSO FA **2018** The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). https://smartlaboratory.org/ravdess/.
31. JYH-SHING ROGER JANG **2021** **Speech and Audio Processing (SAP) Toolbox.** http://mirlab.org/jang/matlab/toolbox/sap.