Formants matching as a robust method for forensic speaker identification

Sergey L. Koval

Speech Technology Center, Saint Petersburg, Russia
koval@speechpro.com

Abstract

"Formants matching" method for robust speaker identification is described and illustrated in practical cases as used in forensic audio. It is a spectral analysis based method which differs from traditional approaches in that it presupposes comparison of articulatory similar events in two compared recordings as opposed to comparison of the same phonemes. Searching for coincidences/differences in uncontrolled movements of speech production organs reflected in higher formants tracks and dynamics makes this method especially robust in situations of noisy audio, different languages, short duration of voice samples.

The method shows high reliability when applied for expert (manual) forensic speaker identification. Its automatic realization gives 1.2% EER for text-dependent voice samples of 3 seconds duration. A similar method is used in the automatic system for speaker identification over phone channels (text and language independent, different communication channels) which shows reliability of 16% EER when both known and unknown speech samples are of 16 sec duration and 8% EER for 96 sec.

1. Introduction

Forensic speaker identification differs from usual speaker recognition in many respects [19-21]. The voice records are often quite short (< 3 sec.); the voice quality is very low, the speech situation is very different, etc. This means that in most cases, the application of automatic identification routines is out of the question. Instead, a computer-aided spectral voice investigation and linguistic speech analysis carried out by an expert is currently used to deal with these cases.

From the beginning of spectral voice investigation in 1949 in Russia [5] up to the present time spectral analysis has remained one of the fundamental stages of decision-making in speaker identity [1,2,6-8,18,19,22]. Both research and practice demonstrate that the main spectral maxima (formants) correlate directly with anatomic and geometrical sizes and structures of speech apparatus, and with acoustic-mechanical properties of its live tissues.

Speech skills and anatomical properties some of which there are controlled and others uncontrolled, or automatic, determine formant positions and behavior [9]. According to the features compared we could classify methods in the following way.

1. Comparison of speech integral properties as a whole [23].
2. Formant comparison of phonetically similar sounds and combinations in comparable contexts. Sometimes such method is called microanalysis [7,8,12].
3. Formant comparison of articulatory similar events [1,2,17].
4. Comparison of spectral structures of the same articulatory dynamics.
5. Comparison of spectral-harmonic structures of laryngeal voice timbre for prosodically similar events.
6. Comparison of spectra and dynamics inside a voice pitch period for comparable phases of vocal folds closure/opening in comparable speech data [18].

Each method has its own advantages, and imposes requirements to the analyzed signal. Usually they are combined in practical usage. The method presented here belongs to section 3 of this classification.

The method suggests indirect comparison of voice tract geometry for articulatory similar acoustical events which is done through comparison of formants [1, 2, 10, 17]. The compared syllables are searched for phonetically equal articulations (equal positions of 2-3 lower formants). For such signal portions the coincidence of consciously uncontrolled high-frequency resonance structures of speech spectra which also have the same low-frequency formants means coincidence of anatomical-geometrical sizes and configurations corresponding to these structures. Having enough of such coincidences for articulatory different sounds, it can be said that accidental coincidence has very low probability, which states the identity or the difference of sizes and subtle geometrical structure of the compared speakers' voice tracts.

2. Method

The method stages are described below, as well as a short discussion and illustrations.

1. General analysis of recordings, choosing speech fragments for comparison. Linear amplitude normalization is necessary. 10-11KHz sampling rate is usually sufficient for identification tasks.
2. Optimizing visual and numerical representation of dynamic spectrograms as to maximum precision and accuracy of formant tracks and correspondence to theory (speech acoustics and phonetics) and experience. Usually an expert chooses adequate type of: spectral analysis type (Fourier-analysis or LPC analysis), spectral precision and spectral model order, scale of spectral, amplitude and time axes, frequency range, length and type of weighting window, analysis inside voice pitch period or averaged at several periods, data normalization and smoothing, visualization type and characteristics, etc.

The typical procedure for enhancing visualization is flattening of average frequency response of a recording (for a whole recording or part-by-part, in more complicated cases), i.e. making frequency response flatter and more standard with the help of inverse filtering (for example as done in SIS or Sound Cleaner Premium software). This procedure usually compensates for the influence of frequency response of the

recording channel and equipment.

3. Listening and preliminary review of dynamic spectrograms, preliminary search for fragments of speech spectrum with an expressive, easily interpreted formant picture with the presence of the 4-th and may be higher formants. The usual minimal unit is a syllable. The poorer quality recording is analyzed in the first place. Formant tracks analysis. Clarifying typical formant behavior in different articulatory situations in order to determine correctly the given speaker's formant tracks in noisy fragments. Getting the spectral structure of the voice for the most opposite phoneme articulation types (A – I – U – E – O). Clarifying the distinction between spectral maxima which are speech formants and stable recording channel characteristics. Getting typical formant tracks in similar articulatory situations and choosing the very typical, repeating formant spectra.
4. Choosing typical basic signal fragments in which typical dynamics of formant tracks of no less than 4 formants is clearly seen and unambiguously determined, which means that spectral maxima are interpreted as formants, each appearance/ disappearance of spectral maxima can be interpreted unambiguously as well. It is necessary to prove that it is indeed the formants that are measured, and not channel or environment characteristics. The formant structure should be typical for a given speaker, that is it enables to verify the speaker at different fragments of his voice sample.
5. Using the basic fragments of the first recording, search in the second recording for the corresponding speech signal fragments, which have matching 1st, 2nd, 3rd formants with the presence of 4th – 5th formants. Those "formants-matched" fragments should have unambiguously interpreted formant structure and coincide in 3 formants with the presence of the 4-5 formants. The phonetic quality of speech sounds pronounced at the moment is not important. The method supposes to take equal momentary articulations (voice tract configurations) in free phonetic context and not equally pronounced phonemes.

Comparison may be done for fragments with disappearance of a formant, if the corresponding spectral maximum movements for the neighboring fragments allow unambiguous interpolation of the disappeared formant track.

Comparison of spectral structures is carried out visually by an expert either in two linked windows of dynamic spectrograms with the help of synchronously moving horizontal cursors, or by comparing spectral frames accumulated with review of an investigated fragment, or by comparing averaged short parts of stationary fragments of spectrograms. Comparison of the 4th, 5th and higher formants for chosen "matched" spectral fragments. In each case: making a decision of their coincidence or difference. The result of comparison is the conclusion about the coincidence/non-coincidence of specific type of formant spectra. If for a specific spectral structure of two compared fragments there is a coincidence of 3 formants and difference in 4th formant, then such a newly found 4-peak spectral structure is taken as a template and a thorough search is done in the second recording for a corresponding new formant vector. Only if it is absent, the conclusion is made that there is a fairly different formant structure.

6. Determining the real preciseness of current formant measurements. Estimating the probability of random coincidence or difference of the obtained spectral fragments "matched" by the first three formants. Estimating the necessary number of coincided or not coincided formant structures of the "matched" fragments to make the decision.
7. Searching for the necessary number of articulatory independent pairs of "matched" spectral fragments. Usually it is enough to find 5-8 types of formant vectors with 5% preciseness of formants measurement. Usually, 18-20 specific sounds are enough.
8. Estimating possibility of the situational factors favorable to changing voice tract geometry and the degree of their influence over formants measurements (objects in the vocal tract, swelling of articulation organs, illness, trauma, extraordinary speaker's position and state, non-standard sound speed in acoustical environment, difference in recording and playback speed for tape recordings or compressed signals, etc.) If the comparable spectral fragments coinciding in three formants in a sufficient number of independent articulations coincide (or differ) in the frequencies of higher formants, then a decision in this type of analysis can be made. Target estimated probability of decision making is usually set not lower than one error at 100 million decisions (approximately, one randomly coincided pair of speakers in 14000 speakers).

Separate attention is to be given to account for differences in the compared recordings due to a different degree of voice nasality.

It is important to remember that apart from vocal tract geometry other factors also influence formants positions. Particularly, any cleavage of air flow in the vocal tract may cause change in the number of formants [11]. Thus, slow opening of the nasal cavity with the same positions of other articulation organs changes the number of formants and gradual (with progressive increase of nasal cavity opening) appearance of nasal resonances and simultaneous displacement or even disappearance of mouth cavity resonances. Figures to illustrate this are taken from [4].

The Figures represent the measured power spectra of Russian vowels [E] and [U] at gradual opening of nasal cavity from pure mouth pronunciation (higher part of figures) up to maximally nasalized pronunciation (lower part of figures). It can be seen that with the increase of nasality, the additional spectral maxima appear: for [E] in range of 700-1200Hz and 2400-2700Hz, and for [U] in range of 500-1000Hz and 2600-3200Hz. Some main "mouth" formants are moved, decrease in amplitude, sometimes disappear. For example, the first 4 formants in these figures have changed in values of 100-300Hz at different degree of nasality.

When analyzing real speech formants, an expert needs to know the tiniest details of this and other articulation phenomena and to account for their influence over formants behavior.

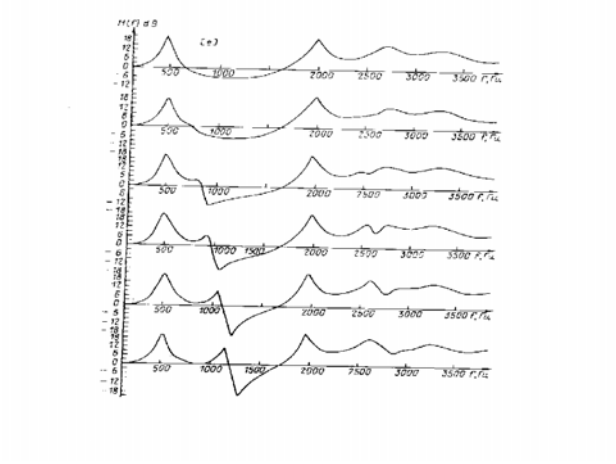

Figure 1: Power spectra of Russian E sound. Degree of nasality gradually grows from up to down of the table.

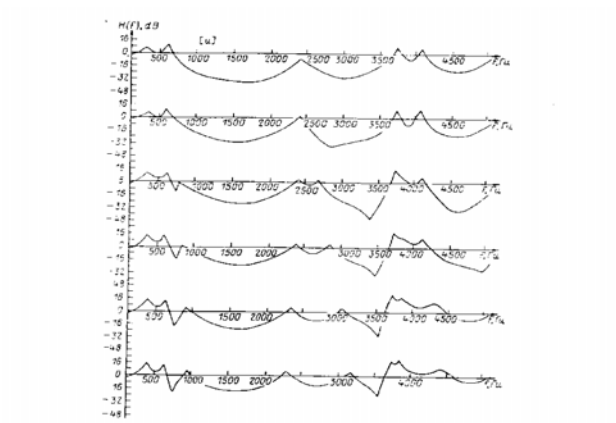

Figure 2: Power spectra of Russian U sound. Degree of nasality gradually grows from up to down of the table.

3. Results and illustration

When the languages of the speakers differ in recordings of the known and unknown voice, a forensic audio expert is to consider the interlingua differences. This situation is often real for bilingual people.

Some established methods of speaker identification are not directly applicable to the multilingual speech. For example, when comparing formants or rhythmical structure of phonetically formally identical syllables, taken from recordings in different languages, there may be some essential differences of one and the same speaker's speech features. Thus, the methods of speaker identification based on comparison of spectral structure of identical phonemes impose special restrictions on the selection of compared speech fragments.

The described method is language-independent, universal and simple in the situation of speech in different languages. Its usage has showed its effectiveness in real forensic audio identification over many years.

The same formal testing procedure was applied. The speech database used for testing is the following: the extracts of microphone speech of 16 Russian speakers, fixed phrases, 5 sessions of 5 different Russian phrases (duration 3-5 seconds), and during one of the sessions 3 phrases in English, with the interval between the sessions not less than 2 weeks. Speakers were native Russians, who studied English for not less than 8 years. All English utterances were quite intelligible as judged by native English speakers with the degree of Russian accent from low to very strong. Recordings were sampled at 16bit 11025Hz.

Speaker identification was done by comparing the English speech of each speaker with the Russian speech of 5 sessions of the same speaker and one Russian session of every other speaker in the database. Total number of 'same-same' comparisons is 80 and 240 of 'same-strange' comparisons. In each compared pair of recordings the expert was looking for 18 basic instant spectra of voiced speech sounds for which any 3 out of 4 lower formants coincide. In these fragments formants should be determined reliably and represent typical formant tracks for different vowels of phonetical triangle, for both recordings.

Particularly, the sounds of A, O, U, E, I types (in Russian) are to be presented (their 1st and 2nd formants should lie inside the typical range of values for the Russian language [13-16]). The Coincidence of formants for every two compared recordings was determined by experts with the necessary preciseness with the help of two horizontal cursors moving synchronously over the frequency axis in two linked windows of the spectrogram. Search for coinciding fragments was done symmetrically in two recordings: fixing formants in the first recording and search of the same formants in the second recording and vice versa. Then for each group of the basic fragments the 4th and 5th formants were analized. For each of sound types, it was accounted for, whether any differences in formant structure are present, whether the basic spectral fragments coincided very well, or if there are no comparable fragments for any of the sound groups in the recordings.

When used by experienced experts for several hundreds real-life identification examinations no wrong decisions were reported. A simpler, purely automatic variant of the method was tested [10] on clear speech: for 100 speakers, with the same test phrases of 3-5 sec duration, 15 sessions within 6 months. 1.2% mistakes was found at EER of false acceptance and false rejection. 8% mistakes was received on the telephone speech database (noisy audio, GSM lines included) [24]. Voice samples at which the mistakes were made were analyzed additionally by experts. Most of the mistakes are due to wrong automatic formants detection. At manual formant detection it is usually possible to escape such mistakes. When the method is used properly by an experienced expert, it is valid for the situation of different languages [17].

Next, the identity/difference decision was made or the impossibility to make a decision was agreed.

Speakers were said to be different in case no less than 3 essentially different basic spectra were found and there were no coincidences for 3 different sound groups. Speakers were said to be the same if no less than 15 coincided base spectra were found and was detected no more than 1 essentially different base spectrum, and for witch there are similar in structure spectra in the second recording.

In the investigated data, the following results were received:

- same-same comparisons: 70 right decisions, 9 rejection of decision, 1 mistake.
- same-strange decision: 207 right decisions, 27 rejections, 6 mistakes.

Later an independent review of the speech data on which mistakes were made was undertaken. It was stated that a more careful and prolonged analysis by an experienced expert eliminates mistakes.

The results listed above show that the formants matching method enables reliable decision-making even in the cases of speech in different languages. The only necessary requirements are a close phonological system and a representative speech material. It is still seen without doubt that only complex examination tools combining aural, linguistic and instrumental methods make a reliable basis for a well-grounded expert's decision to be presented to the court.

4. Conclusion

The formants matching method is based on comparison of formants in articulatory similar events as opposed to the analysis of phonetically similar contexts. This method adds to forensic audio one more tool for instrumental identification procedure. Being rather simple and formal in implementation it may boost further development of automatic speaker identification such as automatic system for speaker identification over phone channels Trawl [25].

5. References

[1] Koval, S., Khitrov, M., Krinov, S. "Formants comparison of similar articulation events for forensic speaker identification", *Proc. of the Workshop Speaker Recognition by Man and Machine: Directions for Forensic Applications*, 1998.

[2] Koval, S., Krinov, S. "Practice of usage of spectral analysis for forensic speaker identification", *Proc. of the Workshop Speaker Recognition and its Commercial and Forensic Applications.* 1998.

[3] Lobanova, M., Raev, A. "Speaker Verification Accounting The Formant Behavior And Phonetic Representation of Enrolled Speech", *Proc. of the Workshop Speaker Recognition and its Commercial and Forensic Applications.* 1998.

[4] Galunov, V., Koval, S., Tampel, I. "Problems of speech production", *Cybernetics issues*. Vol.22, pp. 60-74, 1981. In Russian.

[5] Kopelev, L. "*Slake my sads*",. Slovo Publishers. Moscow. 1991. in Russian.

[6] Kersta, L.G. "Voiceprint Identification", *Nature*, v.196, pp.1253-1257. 1962.

[7] Kuenzel, H.J. "*Sprechererkennung. Grudzuege forensisher Sprachverarbeitung*", Heidelberg: Kriminalistik Verlag, 1987.

[8] Hollien, H. "*The Acoustics of Crime. The New Science of Forensic Phonetics*", Plenum Press. New York. 1990.

[9] Fyodorov, A., Yurgenson, A. "Analysis of speech production by Xray methods" in *Communication equipment*, 1976, TPS series, v. 9, pp. 3-12. Moscow. In Russian.

[10] Koval, S., Labutin, P., Raev, A. "Automatic Speaker Recognition using Formants-Based Nearest-Neighbour Distance Measure", *Proc. EUROSPEECH'95*, 1995, v.2, pp. 341-344.

[11] Sorokin, V. "*Speech production theory*". Moscow: Radio and communication, 1985. In Russian.

[12] Popov, N., Linkov, A., Baycharov, N., Kurachenkova, N., edited by Fesenko, A. "*Person identification with Russian speech recordings with the help of automatic system Dialect*". Moscow. 1996. In Russian.

[13] Bondarko, L. "*Modern Russian Phonetics*". Saint-Petersburg: University Press, 1998. In Russian.

[14] Zlatoustova, L., Potapova, R., Potapov, V., Trunin-Donskoy, V. "G*eneral and applied phonetics*" Moscow: university Press, 1997. In Russian.

[16] Kuznetsov, V. "*Vocalism of speech*". Saint-Petersburg: University Press, 1997. In Russian.

[17] Koval, S., Khitrov, M., "Formants based speaker identification when analyzing recordings in different languages" Proc. Int. Conf. Law enforcement informatization. Moscow, 2003. In Russian.

[18] Rodman, R., McAllister, D., Bitzer, D., Cepeda, L., and Abbitt, P. "Forensic Speaker Identification Based on Spectral Moments". *Forensic Linguistics*, 9(1), pp 22-43, 2002.

[19] Rose, P., "*Forensic Speaker Identification*". London, New York: Taylor & Francis, 2002.

[20] Champod, C., Meuwly, D. "The inference of identity in forensic speaker recognition" *Proc.s of RLA2C*, pp.125-134. 1998.

[21] Braun, A., "Speaker identification - a real challenge for forensic statistics". *Proc. Fifth International Conference on Forensic Statistics*. 2002.

[22] Goldstein, U.G., "Speaker-identifying features based on formant tracks". *J. Acoust. Soc. Am.*, 59(1), pp.176-182, 1975.

[23] Markel, J.D., Davis, S.B. "Long-term feature averaging for speaker recognition". *IEEE Trans. Acoust., Speech, Signal Processing*, 27(1), pp.74-82, 1979.

[24] RUSTEN. Russian Switched Telephone Network database, *STC*, 2003.

[25] TRAWL Authomatic Speaker Identification and Voice Card Comparisons. Announcement. *The Phonetician.* Vol. 89. pp. 82-84. 2004..