TIME STRUCTURE AND DETECTION OF THE MULTIVOICED SEGMENTS IN MIXED SPEECH

Jean-Sylvain Liénard, Claude Barras & François Signol

LIMSI-CNRS, Université Paris XI, Orsay, France
jean-sylvain.lienard@limsi.fr; claude.barras@limsi.fr; signol_francois@yahoo.fr

ABSTRACT

When two speech signals are mixed in a single channel the voiced parts of any of them remain mostly unaltered during the voicing interruptions of the other, i.e. pauses and voiceless consonants. The mixture is made of 3 types of multivoiced segments noted 0V (unvoiced), 1V (one voicing) and 2V (two voicings). A statistical study of read-aloud texts reveals that total time spent in the 1V state is twice as long as the time spent in any of the other states. The HSC multipitch algorithm, based on a specific mechanism that eliminates the f_0 halving and doubling errors, is used to locate the 3 segments types in the signal. This feature is illustrated by the task of spotting a short utterance repeatedly mixed with a long text.

Keywords: voicing, pitch, multipitch, speech separation

1. INTRODUCTION

Today the Pitch Estimation Algorithms for a single speech signal (sPEA) produce more voicing errors than f_0 errors. The most part of them occur at the beginning and end of the voiced segments, where the periodicity is uncertain.

The problem remains and even augments in the f_0 estimation of mixed speech signals (co-channel speech), which is important in the speech separation perspective [1, 3]. In addition to difficulties of the single voice case, spectro-temporal interferences between the voice sources produce unpredictable errors. Determining whether a given frame or segment of the mixed signal comes from one or several periodic sources is a challenging issue, with applications in several domains: speech separation, automatic recognition, diarization, transcription alignment, database annotation.

The first part of the present study deals with the time structure of the voiced and multivoiced segments in the Keele database [6]. The second part deals with their automatic detection in the same data, using a Multiple Pitch Estimation Algorithm (mPEA) of which only the Voiced-Unvoiced decision is taken into account.

2. VOICED AND MULTIVOICED SEGMENTS

In a single-speaker signal voicing may be defined either acoustically as the presence of a "voice bar", or phonologically as a distinctive feature. The terms Voiced and Unvoiced (V and U) used in this paper refer to the acoustical meaning. In a 2-speaker mixture this binary categorization does not suffice because three types of frames or segments may appear, noted 0V for no voicing, 1V for a single voicing, and 2V for two voicings.

2.1. Voicing in single speaker signals

Any part of a single speaker signal may be described acoustically as an alternate series of V and U segments, with some uncertainty on their boundaries.

Table 1 gives for each Keele speaker (5 females, 5 males) the total reading duration of the text "The Northwind and the Sun" (pauses included), as well as the cumulative fraction of time spent in the V and U states. Those figures were computed from the f_0 labels provided by the authors, who used an autocorrelation sPEA working on the associated EGG data (window 25.6 ms, manual corrections).

Table 1: duration (s) and % cumulative duration of the V and U states for the 10 speakers of the Keele database. Global means are 50.3% V and 49.7% U.

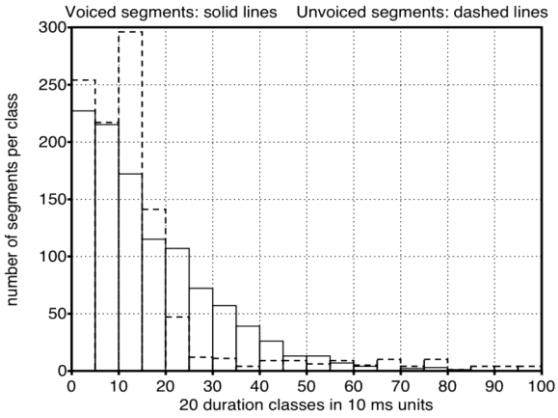
spkr >	fl	f2	f3	f4	f5	m1	m2	m3	m4	m5
dur (s)	32.2	33.7	30.5	31.6	38.7	37.4	31.9	27.2	33.7	40.3
%V	47.5	56.4	49.5	57.1	48.0	48.7	43.4	53.8	48.2	51.4
%U	52.5	43.6	50.5	42.9	52.0	51.3	56.6	46.2	51.8	48.6

The above results suggest that the V and U states have about the same 50% probability. They contrast with those reported in [4], i.e. approximately 75% V and 25% U. This apparent discrepancy reflects the fact that the mentioned study was built on a phonological basis (counting the voiced versus unvoiced phonemes from a

phonetic transcription), while the present study deals with acoustical measures and takes the pauses into account.

Figure 1 shows the duration distribution of both types, all speakers pooled. The mean duration is the same: 156 ms for the V segments, 158 ms for the U segments. But the distributions differ in some respects. The frequency of occurrence of the V segments decreases regularly as their duration increases, up to a maximum of 800 ms. On the other hand, the distribution of the U segments exhibits a clear mode around 120 ms and falls rapidly until a value of some 250 ms is reached. Then one observes pauses, less and less frequent as their duration increases up to 1.7 s.

Figure 1: Histogram of the V and U segments durations from the whole Keele database.



The 120 ms mode of the U segments may reflect the voicing interruption of the voiceless consonants (phonemic value), while the long interruptions (pauses, above 250 ms), contribute to the prosodic/semantic structuration of the discourse.

2.2. Voicing in mixed signals

When two single-speaker sequences are mixed, the multivoiced segments 0V, 1V and 2V exhibit a different distribution. Table 2 gives their cumulative and mean durations, for all the database files concatenated and superimposed to themselves in a different order (counts from the given f_0 labels).

Table 2: % duration for the 3 multivoicing states in a random mixing of the Keele database with itself.

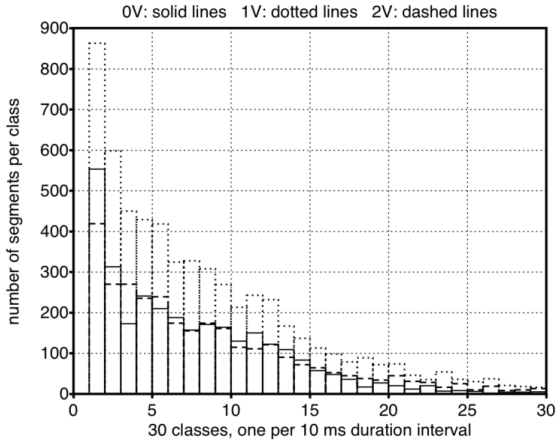
voicing status	cumulative duration (%)	mean duration (ms)
0V	25.4	80
1V	49.0	85
2V	25.6	82

The observed cumulative durations could be predicted by simple calculus, knowing that the V and U states of the constituents were equally probable. The main point is that the 2V segments occupy only 25% of the total duration. If the cumulative durations of the V and U single-speaker segments were those reported in [4], the cumulative durations of the 0V, 1V and 2V segments in the 2-speaker mixture would be 6%, 38% and 56%, respectively, i.e. the 2V segments would by far outnumber the other segments.

A consequence of this finding is that a separation strategy, based in the first place on the detection and identification of the sole 1V segments, may be envisioned. This view relates closely to the notion of "glimpsing" [2].

Figure 2 shows that the distributions of the 3 types of segments are very much alike: similar mean and max values (about 80 ms and 300 ms), and similar smoothly decreasing shape. The 25/50/25% proportion for the 0V/1V/2V segments keeps approximately true for all duration classes.

Figure 2: distribution of the 0V, 1V et 2V segments durations in the mixed signal formed by addition of two permutations of the Keele database.



3. MULTIPITCH ESTIMATION

Few mPEAs are available today. The one used here works in the frequency domain and is called HSC (Harmonic Suppression Comb) [5, 7].

3.1. Principle of the HSC multipitch algorithm

HSC strictly performs a frame-to-frame analysis, without any top-down knowledge or post-processing. It makes use of spectral combs, i.e. sets of discrete unit values (teeth), multiples of the base frequency f_c that covers the frequency interval given for f_0. The crosscorrelation with the modulus

of the short-term spectrum yields a "pitch function" which has a series of peaks noted (p,q) where $f_c/f_0=p/q$ (p and q positive integers). The peak corresponding to the fundamental is the (1,1) peak. There may be several (1,1) peaks if several periodicities are present.

In a single-pitch estimator the main peak (1,1) is usually larger than the parasitic peaks, so that a simple maximum detection performs correctly. But in a multipitch situation the main peak of the second sound may be smaller than a parasitic peak of the dominant sound, which causes a harmonic or sub-harmonic error.

The specific feature of HSC is a mechanism that suppresses most of the parasitic peaks by jointly using two families of irregular combs: the missing teeth combs eliminate the sub-harmonic errors and the negative teeth combs eliminate the harmonic errors. This feature provides a crucial advantage in detecting the voicings in mixed speech.

3.2. Performance of single-pitch HSC on the Keele single-speaker data

In order to test the mPEA in the usual single-speaker situation HSC was allowed to deliver only one pitch/voicing hypothesis per frame. The results were compared to the pitch labels provided with the database (table 3).

Table 3: HSC monopitch error rates (in %) on the Keele single-speaker data: V>U (undervoicing); U>V (overvoicing); total Voicing Error Rate (VER); pitch Gross Error Rate (GER) at 20% max deviation.

Error rates %	V>U	U>V	VER	GER
ref= Keele data	2.4	2.3	4.7	0.9

The voicing errors were counted on the set of valid frames, i.e. the frames where the pitch could be correctly defined according to the pitch labels and the chosen f_0 interval (75-600 Hz). The pitch errors were counted on the set of frames jointly declared voiced by HSC and the pitch labels.

The above usual pitch/voicing measurements do not easily extend to the multipitch case. In the rest of the paper the Recall/Precision measurement is preferred. The Recall indicates to what extent all of the correct information has been recovered, while the Precision tells what proportion of the recovered information is correct. Those quantities are given as percentages, as well as the F-measure which is a tradeoff (harmonic mean) between recall and precision. Table 4 shows those figures for the above voicing evaluation.

Table 4: Voicing measurements of table 3 expressed as Recall/Precision/F-measure.

Voicing detection %	Recall	Precision	F-measure
ref= Keele data	95.3	95.5	95.4

Those measurements indicate that the voicing detection by HSC is in good agreement with the database labels, despite the differences in the analysis methods used. In the next section the HSC voicing estimates on single signals are considered as the reference, in order to fairly evaluate the multipitch behavior of the algorithm.

4. MULTIPLE VOICING DETECTION

4.1. 2-voicing estimation of a speech mixture

The speech material consisted of a concatenation of the 10 Keele speech files, after level equalization. This operation was performed twice, in different permutation orders, yielding two signals having the same total duration. The voicing references were given by HSC in monopitch on both signals. A "reference multivoicing" function was established by adding the number of voicings found in both, thus taking the value 0, 1 or 2.

Then the two signals were mixed and the mixture was analyzed by the same HSC, now in bipitch mode (2 voicing hypotheses), yielding a "detected multivoicing" function varying in the same range. Both functions were compared using the recall/precision measurements applied to 4 types of entities: voicing hypotheses, 0V, 1V and 2V frames (table 5).

Table 5: Multivoicing results of HSC Multipitch on a mixture of 2 different permutations of the 10 Keele sequences.

Voicing detection %	Recall	Precision	F-measure
cumulative Voiced hypotheses	94.3	83.1	88.3
cumulative 0V frames	97.1	98.0	97.6
cumulative 1V frames	62.0	84.2	71.4
cumulative 2V frames	79.1	52.0	62.8

The first line, compared to table 4 (single-pitch estimation on single-speaker signals), demonstrates the good behavior of HSC in this bipitch estimation of a 2-speaker mixture: the degradation, in terms of individual voicings, is limited to a small loss in recall (1% more misses) and a moderate loss in precision (12.4% more false alarms). The next 3 lines, in terms of multivoiced frames, indicate that:

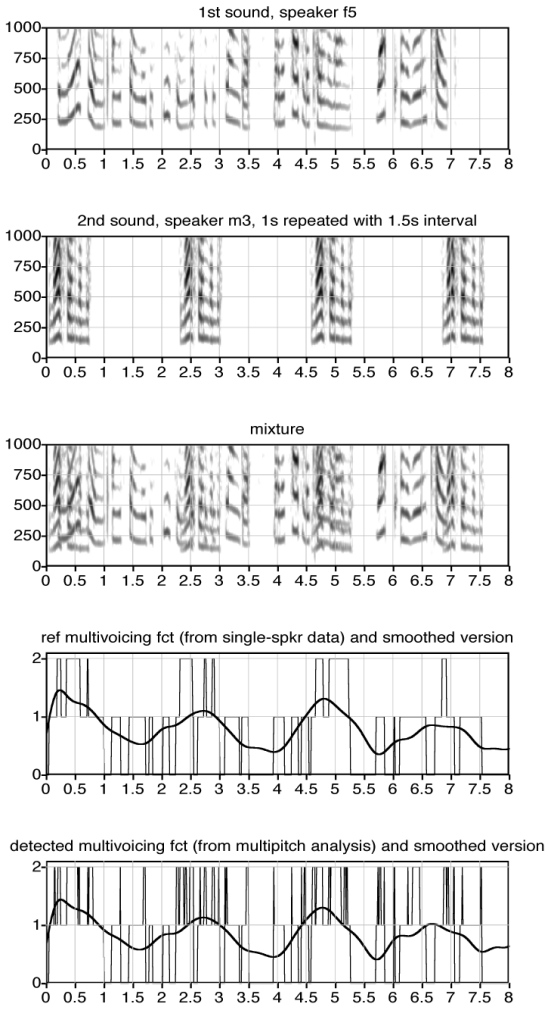
- the 0V frames are well detected

- the 1V frames are poorly recovered (many misses) but with a good precision (few false alarms).
- the 2V frames are well recovered but with a poor precision (many false alarms).

5. LOCATING AREPEATED SHORT UTTERANCE IN A SPEECH MIXTURE

We are now interested in locating a short utterance repeatedly mixed with a long speech signal. It may happen that two mixed signals do not locally produce many 2V frames, because some voiced frames of the first speaker appear at instants where the second is in the unvoiced state. If the sequences are long enough the distributions of figure 2 reappear, but there is a possibility that short sequences cannot be detected on the sole basis of the 2V voicing.

Figure 3: from top to bottom: spectrograms (8s) of the 1st sound, the second (repeated), the mixture, the reference multivoicing function and its smoothed version (thick line), the detected multivoicing function and its smoothed version (thick line).



In order to illustrate this point a Keele sequence (speaker f5) has been repeatedly mixed with the first 0.8 s of another sequence (speaker m3, 5 syllables) at 1.5 s intervals. The beginning of the mixture (8 s) is illustrated in figure 3: spectrograms of the constituents, the frame to frame multivoicing functions and their smoothed version (1 s averaging and low-pass filtering). In this example the first 3 occurrences of the short sentence could be detected with a threshold value of 1 or 0.9; the 4th would be missed because there is no real overlap of the signals in that region.

However the reference and test smoothed functions are very similar. This means that despite the false alarms observed in the 2V detection, the information provided by the multivoiced segments remains and could help to locate the superimposed sequences.

6. CONCLUSION

The mixture of two speech signals in the reading style can be acoustically analyzed as a series of segments pertaining to 3 categories according to the number of simultaneous voicings: 0V, 1V and 2V. The 1V segments occupy about half of the total duration. Using a multipitch estimation algorithm as a frame-to-frame multivoicing detector gave satisfactory results. Time integration of the number of detected voicings looks promising for spotting short speech sequences superimposed to a long speech flow.

7. REFERENCES

[1] de Cheveigné, A. 2005. Multiple F0 estimation. In: Wang, D.L., Brown, G.J. (eds.), *Computational Auditory Scene Analysis*. John Wiley and Sons.

[2] Cooke, M.P. 2003. Glimpsing speech. *J. of Phonetics* 31 (3-4), 579-584.

[3] Divenyi, P. (ed.). 2005. *Speech Separation by Humans and Machines*. Dordrecht, Kluwer Ac. Pub.

[4] Hu, G., Wang, D. 2008. Segregation of unvoiced speech from nonspeech interference. *J. Acoust. Soc. Am.* 124, 1306-1319.

[5] Lienard J.S., Barras C., Signol F. 2008. Using sets of combs to control pitch estimation errors. *Acoustics'08*, published in ASA *Proc. Of Meetings on Acoustics*, 4(1).

[6] Plante, F., Meyer, G.F., Ainsworth, W.A. 1995. A pitch extraction reference database. *Eurospeech95*, Madrid, 837-840.

[7] Signol, F. 2009. *Evaluation de Fréquences Fondamentales Multiples en vue de la Séparation de Parole [...]*. Doctoral thesis, Paris XI University, Orsay.