CONVERTING SPEECH SIGNALS TO PHONOLOGICAL FEATURES

Henning Reetz
Department of Linguistics, University of Konstanz, Germany

ABSTRACT

The FUL speech recognition system [1] uses underspecified phonological features for the lexical access of words. These features are extracted from the acoustic signal by standard LPC-analysis and by the computation of rough spectral shape parameters. The obtained streams of formants and spectral shape parameters are corrected on a local and global basis by an algorithm that mimics the visual perception of the formant/shape traces. The corrected traces are converted by simple logical decisions into sets of features. Short disturbances are removed from these feature sets and the individual features are time-aligned to give sets of synchronously changing features. These aligned feature sets are used to access possible word candidates from a lexicon of 50 000 base word forms.

1. THEORY

Most automatic speech recognition systems use spectral templates of short-time windowed speech signal to compare the incoming speech signal with sequences of templates for stored words. The identification of sequences of computed templates with stored template sequences of words is usually performed with a Hidden Markov Model. These models are trained with large sets of acoustic material. Any change of the acoustic characteristics that influences the spectral templates has to be corrected. Consequently, different microphone or line conditions, different speakers or dialectal variants require an adjustment of the system. This adjustment is normally performed by a training session or an online estimation for the corrections the system has to make.

The FUL system operates on a completely different principle. No spectral templates are computed from the speech signal to access the lexicon, nor is the signal analysed in great detail for acoustic evidence of individual segments and their boundaries. Instead, the signal is converted into speaker independent sets of phonological features. These features are compared with the feature sets stored in the lexicon using a ternary logic [1]. The task of the acoustic front-end described here is (a) to remove linguistically irrelevant information, (b) to use speaker independent acoustic characteristics to compute the features, and (c) not to exclude potential word candidates due to computational faults or poor signal quality. The general design principle of the system is to use simple and only rough measures that through cooperation form a stable system. This behaviour could be compared to beavers building a dam – they use many roughly cut stems of trees und combine them in a way that forms a robust dam. The idea is to be able to handle massively faulty data rather than to expect clean data with a high signal-to-noise ratio.

2. SYSTEM DESCRIPTION

The conversion of the speech signal to phonological features is performed in two steps with one correction applied in each step.

First, the signal undergoes a spectral analysis that delivers LPC formants and some rough spectral shape parameters. A heuristic filter is applied to this computed sequence of formant and spectral shape data to convert the raw (formant) data into (formant) tracks as they would be drawn by a human investigator of the visually presented data. In the second step the formant and spectral shape data are converted by simple heuristic filters into phonological features. The feature data again undergoes a heuristic filter to remove irregularities in the stream of features. Additionally, the individual tracks of features are time aligned so that changes in the feature sets occur in all features at the same time. These steps and the corrections are described below.

2.1. Spectral analysis

A standard LPC analysis and a standard FFT power spectrum are computed from the speech signal from 20 ms Hamming windowed stretches of speech. The first five formants are taken from the LPC analysis, the overall spectral slope, the spectral slope below, and the slope above 2.5 kHz are computed from

the FFT spectra. The window step rate is 1 ms. That is, in the time domain a very dense sequence of quite roughly computed spectra is available. The intention is to convert the 'online' speech signal stream into an online stream of spectral data. This temporal density might not be necessary for the system at every part of the signal but it is currently used as redundant information to separate accidental glitches in the data from systematic variations due to properties of the signal.

The outcome of the first step, the spectral analysis, is a set of 5 formants and 3 spectral slope parameters computed every millisecond from a 20 ms window.

2.2. Spectral correction

The spectral analysis is performed for each window individually without considering the preceding and following windows. The lowest five roots of the autocorrelation function or the first five maxima in the LPC spectrum are numbered from 1 to 5. (Here I describe only the correction procedure for formants but the algorithm operates similar for the spectral slope parameters. Furthermore, the algorithm is described as if the signal is stored in the computer; the algorithm can operate online but the description becomes rather awkward). If a formant is 'missing' in one or more consecutive windows, either because it is not part of the acoustic signal, because it is camouflaged by a zero or merged with an adjacent formant, or because the spectral analysis failed to register it, the numbering of the formants changes whereas the formant tracks are continuing. There is no way to absolutely decide what is the 'correct' number of a formant but in general a human, looking at formant tracks, will identify stretches of formant tracks and can label them. The correction step described here tries to mimic this behaviour.

For each formant of a given frame (i.e., the set of formants computed from a windowed signal part) the directly left and right adjacent frames (i.e. the formants of the preceding and following 1 ms) are investigated. Each adjacent formant that is less than 50 Hz away and does not have another formant within a range of 200 Hz is taken as a direct neighbour of a formant. If a direct neighbour cannot be found on either the left or the right of a formant, the next but one and the next but two frames are investigated. If a neighbouring formant can be found there formants in the intermediate frames are generated by linear interpolation. This first operation of the formant correction algorithm leads to stretches of connected formant tracks that cannot cross each other. In a second pass, those formants are investigated that have two adjacent formants within 200 Hz. In these cases, the determination of the sequence of formants and the interpolation procedure ensures that two formant tracks do not cross. The first two passes link formants according to their temporal behaviour. A third pass investigates the formants in terms of their similarity within one frame. Formants are connected to their adjacent tracks if the frequencies of all formants of a frame are shifted relative to the preceding or following frames by a similar amount. This operation ensures that formant shifts due to disruption of the speech signal, e.g. at a plosion, are treated as sporadic formant values. The fact that *all* formants of a frame change in a similar way is used to treat them as formants that bear some information. After this procedure, all singular formants or formant tracks that are shorter than 5 ms are removed in a fourth pass. Finally, the formant tracks are numbered from low to high frequency with 1 to 5 with the constraint that formant 1 must be below 1 kHz. If the first formant track is above 1 kHz the counting of formants starts with '2'.

2.3. Converting spectral parameters to phonological features

The corrected spectral parameters are used to compute phonological features. These features should reflect important properties of the speech signal and they should be relatively independent of the speaker and acoustic line properties. It is *not* intended that the features exactly match the features that are stored in the lexicon; only features that are impossible for a certain sound (e.g. a low vowel should never be high) should not be computed from the signal. That is, a low vowel should preferably have the feature [low] in the acoustic signal, but a missing low computed from the signal does not exclude it from the subsequent processing. Only if the feature [high] is computed from the signal a low vowel is excluded in the lexical access.

The system uses very broad acoustic characteristics to define the phonological features. We investigated two databases [2, 3] and found that it is possible to characterise

individual segments for individual speakers with more complex acoustic characterisations, but these characteristics could not be extended to different speakers. We also tested some speaker normalisation procedures but we found a rather broad specification of acoustic characteristics for the phonological features for the raw spectral data more efficient than the application of a speaker normalisation procedure prior to the classification. We observed an improvement by the speaker normalisation procedures with carefully measured formants of speech signals recorded under optimal conditions, but we observed a deterioration with the normalisation procedure with automatically extracted formant data (see [4] for detail).

Most of the 13 phonological features we use (consonantal, vocalic, continuant, RTR, voice, strident, abrupt, nasal, labial, coronal, dorsal, high, low) are defined by simple acoustic characteristics. For example, the feature [high] is defined by F1 < 450 Hz. These very simple characteristics are often true for sounds that do not belong to a particular feature (e.g., some other sound that is not [high] might have an F1 below 450 Hz) but crucially, this acoustic characteristic seldom misses any sound that is [high]. That is, the acoustic characteristics are chosen so that all members of a particular feature are captured, and other sounds might be included as well, but no member is missed. The construction of the matching process and the lexicon (see [1]) eliminates implausible candidates later. The rationale behind this very relaxed procedure is that in running speech a speaker can deviate from any 'norm' of acoustic characteristics of a sound due to assimilation, coarticulation, dialect, vocal tract parameters, and others. The FUL system does not have such a 'norm' (for example, as it is set up during the training phase of a Hidden Markov Model). The system only expects that the feature [high] is acoustically characterised by a low first formant and that a high first formant would be identified by a listener as a low vowel.

Certain characteristics can be more complex. For instance, the feature abrupt includes an investigation of the change of all formants in a short period of time. The acoustic characterisation of this feature includes a description of the temporal development of all formants, and not simply the sudden increase of energy, as might be expected from a plosion release. The reason is that stops in running speech are not necessarily marked by a nice closure and a sudden release, as it is common in laboratory speech. Stops can become quite reduced to a short disturbance in a speech signal without a preceding stretch of a pause.

The outcome of the extraction of features from the spectral data of the speech signal is a stream of phonological features. Like the spectral data, the features are computed every millisecond. Unlike the spectral parameters some features can include a relation to adjacent spectral data, as discussed for the feature abrupt, which by definition includes the inspection of a change over time.

2.4. Correcting phonological features

The phonological features computed in §2.2 are described by acoustic characteristics. Some features inspect the time course of the signal (or its spectral data) but most features are set (or not set) on the instantaneous presence of their acoustic characterisation. Consequently, features can be present or not at any time and they can change their state independent of other features. This can lead to the existence or non-existence of features for a time period of only one millisecond, which can be due to a noisy signal channel or a fault of the formant or spectral shape computation. It can also be a consequence of the speaker's behaviour. For example, the change from one sound to another sound normally changes more than one feature. The speaker does not change all these features at the same millisecond. Rather, by the nature or articulation, within a certain period of time, feature change one after the other in a probably rather disorganized manner. That is, there is a transition state between the production of two sounds. Actually, some models of automatic speech recognition focus on these periods of changes in the signal [5] but the FUL system tries to locate more stable areas. The reason for this approach is based on the consideration that any noise will lead to many sudden changes in the signal and that the more robust parts in a speech signal must be more salient to be distinguishable from any noise. Furthermore, the entries in a lexicon are defined by a sequence of feature sets. The transition state would be interpreted as a sequence of rapidly changing feature sets. If they would be forwarded to the lexicon, the lexicon would have to store all combinations of transitional

states, exactly the opposite of the FUL approach, which wants to keep variation out of the lexicon and handle variation in the matching process during lexical access.

This overview described the task of the correction system for the phonological features: first, spurious changes have to be removed and, second, changing features have to be synchronised. A special treatment is required for the features [abrupt] because this feature is (a) defined by a very rapid change, which is for all other features considered to be an error condition, and (b) can have a very brief period of acoustic characteristics to define the place of articulation adjacent to it, which would for all other features considered to be a spurious occurrence. Consequently, the existence of the feature [abrupt] forced a special treatment of all features in the adjacent ±20 ms that is not described in detail here, and it blocks all operations in the other feature tracks that are described below.

For all other features, the algorithm first searches each feature track for gaps that is shorter than 5 ms and fills such gaps by inserting the feature. Second, isolated stretches of features in a track that are shorter than 15 ms are removed. These two corrections are the only corrections currently applied to the feature tracks. More specific operations on individual features are under test. These 'improved' filters seem to operate well for individual speakers, but applied to all speakers in the same way they seem to decrease the overall performance of the system.

The synchronisation of the individual features is first performed for the feature [abrupt]. All features directly adjacent to the feature [abrupt] that persist for at least 10 ms are set together with the feature abrupt for 10 ms, otherwise they are removed. Then, the procedure described below is applied without investigating the other features outside these 10 ms 'abrupt' windows.

If any of the features changes its value (i.e., it becomes existent or it disappears) then the next 20 ms are investigated. If no other feature change occurs, the 20 ms are treated as a set of stable features, that are used to access the lexicon. If any feature changes its value during these 20 ms, than the period is considered as unstable and a new 20 ms search for stability is triggered. Roughly speaking, this procedure leads to sets of features that are at least 20 ms long, but that raise only one lexical access. The length of the stability of a feature set could be used as information as well, but in the present implementation this information is disregarded.

SUMMARY

The acoustic front-end of the FUL speech recognition system converts the speech signal into a set of spectral parameters. These parameters are corrected to reduce the influence of poor signal quality or estimation faults of the applied algorithms. The spectral parameters are then converted into a set of phonological features. Spurious gaps or occurrences of features are removed and the individual features are synchronised. Only 'stable' feature sets are used to access the lexicon.

Ongoing work investigates alternative methods in all these steps. The present results indicate that formants are very robust parameters for many sounds but that other parameters than the overall spectral shape might be needed to describe all phonological features reliably. The extraction of spectral parameters is tested with methods that improve the estimation of the parameters in the presence of background noise, including simultaneous speakers. Speaker normalisation procedures seem not to improve the system. The conversion from the spectral parameters into phonological features is also tested with algorithms that take more than one feature into account. The interconnection of features can improve the reliability of the decisions but it can also transfer error from one feature onto other features. Therefore, we try to keep the dependencies between the definitions of the features to a minimum. In the same vain, we want to continue to use simple and rather independent criteria to filter and synchronise the individual feature tracks.

REFERENCES

[1] Lahiri, A. 1999. Speech recognition with phonological features. *The XIVth International Congress of Phonetic Sciences*, San Francisco.

[2] IPDS (1995). *The Kiel Corpus of Spontaneous Speech, Vol. 1*. CD-ROM. Kiel: Institut für Phonetik und digitale Sprachverarbeitung.

[3] TIMIT (1993) *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. CD-ROM. LDC.

[4] Reetz, H. (1999) Evaluation of spectral estimation techniques. *Eurospeech 99*, Budapest.

[5] Stevens, K. N., Perkell, J. S., & Shattuck-Hufnagel, S. (1997). Speech Communication. *MIT-Research Laboratory for Electronics Progress Report*, *140*, 353-367.